Fat and Fatter
Monthly Crash Risk and Investor Trading[*]

QIAN YANG[†]

First Draft: January 6, 2021
This Version: April 9, 2021

Abstract

I predict monthly ex-ante crash probabilities and jackpot probabilities via novel machine learning methodologies. Using the predicted crash probabilities as a proxy for monthly crash risk, I show that the risk predicts negative return spread in both portfolio tests and cross-sectional tests. Institutional and retail investors tend to buy high crash risk stocks, rendering them overpriced, and predicting a negative return spread subsequently. Using Robinhood introduction of commission-free option trading at the end of 2017 as a quasi-experimental setting, together with textual information from Reddit, I show that retail participation significantly increases ex-ante stock crash risk, and this effect is stronger for smaller firms.

Keywords: Crash Risk, Cross-Section of Stock Returns, Imbalanced Learning, Machine Learning, Robinhood, Tail Risk, Wallstreetbets.

[*] I thank my advisors Naveen Khanna and Hao Jiang for their support and encouragement. I thank Kirt Butler, Nuri Ersahin, Ryan Israelsen, Dmitriy Muravyev, Mark Schroder, Andrei Simonov, Parth Venkat, Hayong Yun, Morad Zekhnini, and participants in MSU brownbag for their valuable comments.

[†] yangqia8@msu.edu, Eli Broad School of Business, Michigan State University.

1. Introduction

In the year 2020, during the raging COVID-19 pandemic, the stock market experienced a dramatic downturn that were followed up by a fantastic comeback. Media attributed the crash and recovery partly to the speculative behavior of retail investors such as the so-called "Robinhood Traders".[1] There are substantial evidence that retail traders are particularly active with the emergence of commission free trading. It is perceivable that their participation should cause an increase in volatility, if we assume that they are noise traders.[2] Then a pertinent question is: do retail investors have the marginal power to increase the probability of tail events?

This paper seeks to answer this question by first developing a model to predict monthly stock-level crash and jackpot probabilities. Using the probabilities as a proxy for tail risk, I show that monthly crash risk robustly predicts negative returns in the subsequent month. By using Robinhood introduction of commission free option trading as a quasi-natural experiment, I show that increased participation of retail investors contribute significantly to stock crash risk.

There has been continuous efforts in predicting ex ante crash probabilities[3]. The literature usually defines crash risk as the probability of stock crash in the next year, which is not suitable for the analysis of investors' short-term behaviors. Furthermore, the typical econometric toolkit the literature has employed is insufficient in forecasting tasks. One important issue is that "crashes" and "jackpots" are extreme events with very low probability of occurrence. This creates a severe "imbalanced sample" problem, where some categories have far smaller sample sizes as compared to others. Using generic classification models like logistic regression will tilt the gravity towards the majority categories, and hence causes

[1] See for example: "When the Stock Market Is Too Much Fun", by *Jason Zweig*, *Wall Street Journal*, December 11, 2020, `https://www.wsj.com/articles/when-the-stock-market-is-too-much-fun-11607705516?mod=searchresults_pos14&page=1`.

[2] See, for example, Foucault, Sraer, and Thesmar (2011).

[3] See for example Conrad, Kapadia, and Xing (2014) and Jang and Kang (2019).

severe bias in the coefficients[4]. It follows that the measurement error for the tail risks can be substantial.

To solve this problem, I introduce Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, and Kegelmeyer (2002)) to balance the sample, and then use logistic Ridge regression to tune and regularize the model, in order to improve the out-of-sample performance. I show that, using this procedure can achieve substantial improvement over the base model with respect to performance metrics for both crashes and jackpots. The F1-scores for crashes and jackpots show meaningful improvement over the base model. Moreover, the crashes and jackpots are sufficiently separated: the unconditional correlation between crash and jackpot probabilities is estimated to be around -25%.

With improved monthly crash probabilities at hand, I show that a zero-cost strategy long in high-decile crash risk portfolio and short in low-decile crash risk portfolio produces consistent and significant negative alpha benchmarking against various asset pricing models, with average alpha of around -1% monthly at less than 1% statistical significance. To show that this risk has incremental power in predicting returns, I run Fama-MacBeth cross-sectional regressions, controlling for a plethora of conventional and anomaly characteristics. This exercise shows that crash risk is consistently and negatively priced.

Next, I examine both institutional and retail investors' trading behaviors with respect to the two tails. It is often assumed and shown to some extent that institutional investors are rational speculators, while retail investors are noise traders. However, it is an open question as to whether these investors are able to distinguish between the left and right tails. If institutions are rational speculators, they should be able to anticipate imminent crashes and jackpots, and earn superior returns. On the other hand, if retail investors are noise traders, their trading could only increase noise, and thus exacerbate stock crash risk. However, to establish causality, we need a natural experiment.

To answer these questions, I use 13F data and Robintrack[5] data to explore the relationship

[4] See for example (Haixiang, Yijing, Shang, Mingyun, Yuanyue, and Bing (2017)).

[5] Robintrack: `https://robintrack.net/`

between crash risk and institutional and retail investor trading. I found that on average, Institutional and retail investors seem to chase the left tail, rendering the high risk stocks overpriced, and revealing a negative return spread subsequently. Then I explore a quasi-natural experimental setting where Robinhood introduced commission-free option trading at the end of 2017. I use textual information from Reddit Wallstreetbets to identify treatment stocks that retail investors participated in option trading after the event, to examine the possibility of information transmission from option trading to stock trading. Through a difference-in-difference analysis, I show that on average, this event significantly increases stock monthly crash probability. And I show that, after the event, investors are faced with higher far out-of-money option prices, increased stock trading volume, and increased total volatility. Moreover, the effects are stronger for smaller firms, but weaker for large firms.

There are several unique contributions this study makes to the literature: first, to the best of my knowledge, this is the first study that jointly estimate short-term one-month ahead crash and jackpot probabilities, and shows that crash risk is robustly priced in the cross-section; second, this is the first study in economics that introduces imbalanced learning methodologies to improve forecasting performance for rare events or the so called "imbalanced sample" tasks to reduce the measurement error; third, this is the first study that utilizes a quasi-natural experimental setting to study the positive impact of retail trading on stock crash risk.

The paper is organized as follows: Section 2 conducts literature review on crash risk and discusses the limitations of the prior studies; Section 3 provides summary statistics for the data used in this study; Section 4 reports the prediction method and results for predicting near-term crash risk; Section 5 shows the asset pricing tests for crash risk for the cross-section of stocks; Section 6 conducts analysis on investor behaviors and their impact on crash risk; Section 7 conducts robustness tests; Section 8 concludes.

2. Literature Review

The literature on crash risk is extensive in both corporate finance and asset pricing. On the corporate side, the literature is mostly concerned with the determinants of firm crash risk. These determinants are often motivated by managers hoarding bad news (Jin and Myers (2006)). The idea is that the hoarding delays the information transmission such that when it is ultimately released, there is a sudden drop of price corresponding to the size of the cumulative bad news. Motivated by this theory, the literature has proposed a list of determinants that could endogenously influence crash risk: earnings management (Hutton, Marcus, and Tehranian (2009)); tax avoidance (Kim, Li, and Zhang (2011)); annual report readability (Li (2008)); CSR (Kim, Li, and Li (2014)); liquidity (Chang, Chen, and Zolotoy (2016)); short interest (Callen and Fang (2015)); governance (Andreou, Antoniou, Horton, and Louca (2016), An and Zhang (2013)).

On the asset pricing side, there is vast literature in option pricing that tries to extract information from options to determine the size of tail risk. Bates (1991) was among the early papers that study the relationship between option prices and crashes. They show that the 1987 stock market crash can be predicted by the unusually high prices of out-of-money S&P 500 futures put options. Further more, the paper indicates that the jump diffusion parameters implied by the option prices show that the crash could be expected. Pan (2002) provide theoretical support for the jump-risk premia implied by near-the-money short-dated options that help explain volatility smirk. Xing, Zhang, and Zhao (2010) study the relationship between implied volatility smirks and the cross-section of stock returns. They show that the difference between implied volatility of out-of-money put options and at-the-money call options show strong predicting power for future stock returns. Yan (2011) show that jump size proxied by the slop of volatility smile predicts cross-section of stock returns. More recently, Barro and Liao (2020) build a new theoretical model for option pricing that links the relative price of far-out-of-money put options with the probability of rare disasters. They show that the relative price of far-out-of-money put options are positively associated with the probability of rare disasters, which they infer from monthly fixed effects in empirical test.

Another direction attempts to directly predict the probability of rare events such as crashes. Chen, Hong, and Stein (2001) use cross-sectional regressions to forecast skewness of daily stock returns. They show that negative skewness can be predicted by recent increase in trading volume and positive returns. Campbell, Hilscher, and Szilagyi (2008) use a dynamic logit model to predict distress probabilities for cross-section of firms. They show that high-distress-risk stocks suffer from lower subsequent returns. Conrad et al. (2014) show that high-distress-risk stocks are also likely to become jackpots. They use a logit model to predict the probability of deaths and jackpots. Further, they find that institutions tend to hold less higher-default and jackpot probability stocks. Most recently, Jang and Kang (2019) exploit a multinomial logit model to jointly predict probabilities of crashes and jackpots using a plethora of predicting variables. They show that institutions appear to ride the bubble instead of trading against high crash risk stocks, and overpricing cannot be fully explained by investor sentiment.

These studies typically examine the probability of crash in the next year. For example, Jang and Kang (2019) and earlier papers define crashes as less than -70% log return for the coming year. Since there is no certainty when the majority of crashes happen in which month, together with the fact that the study uses Fama-French three-factor model (Fama and French (1993)) plus a momentum factor (Carhart (1997)), it is tenuous to argue that this probability can predict monthly stock returns. Indeed, using more recent sample period from 1996 to 2019, a replication of these results while benchmarking against newer asset pricing models fails to produce convincing negative alphas. As I show in Appendix A.1, while benchmarking against CAPM, Fama-French three-factor, and momentum augmented four-factor models, the zero-cost high-minus-low crash risk portfolios show significantly negative alpha, the alphas quickly turn economically and statistically insignificant when the five-factor model (Fama and French (2015)) is used.

The second issue that the literature has yet to address is a severe imbalanced sample problem: by definition, tail probabilities are very low compared to normal conditions, where the crashes and jackpots are extremely rare. As noted in Ripley (1996) and King and Zeng (2001), the poor finite sample properties in the imbalanced sample context would bias the coefficients, as the majority class will be much better estimated than the minority class. This then casts doubt on whether we can confidently use the predicted probabilities as a valid proxy for ex ante crash risk.

There is abundant literature on the relationship between investor trading and stock returns, volatilities, and potentially higher moments. For example, De Long, Shleifer, Summers, and Waldmann (1990a), De Long, Shleifer, Summers, and Waldmann (1990b), and Abreu and Brunnermeier (2003) provide the theoretical and empirical evidence of positive feedback traders and their potential impact on market. Greenwood and Nagel (2009) show that inexperienced institutional investors might help the formation of bubbles. On the other hand, it is often assumed by literature that retail investors are by and large “noise traders” that could trade too much (Barber and Odean (2000)), and that those speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar (2013)). Recent evidence from Robinhood traders show that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber, Huang, Odean, and Schwarz (2020)), while there is also evidence that mimicking portfolios based on the characteristics of “Robinhood traders” do not seem to underperform the market, but instead could be a market stabilizing force (Welch (2020)). These seemingly conflicting evidence begs for further studies. In particular, in the present context, the question is whether retail investors can exacerbate stock crash risk through their participation and trading.

Finally, this study is also related to the emerging literature that studies the implications and applications of machine learning models in asset pricing and corporate finance.

3. Data

3.1. Variables

For definition of crashes and jackpots, I use log monthly returns of -20% and 20% as the cutoff points. It is reasonable in the following sense: prior literature uses log annual return of -70% and 70% as the cutoff points; since I estimate near-term monthly crash risk, -20% and 20% are at reasonable magnitude; further more, the unconditional probabilities of stocks reaching these monthly returns are comparable to that of using -70% and 70% in annual returns. Then the dependent variables are defined as categorical, where $crash = -1$, $jackpot = 1$, and $plain = 0$. For independent variables, I use Compustat quarterly data to construct accounting variables, analogous to the annual measures used in Jang and Kang (2019), where I transform the frequency to short-term intervals to match the predicting task. These fundamental variables include: past 3-month market return, past 3-month stock excess return relative to CRSP value-weighted market return, book-to-market ratio, asset growth, return on equity, total volatility, total skewness, size, detrended turnover, firm age, tangibility, and sales growth. On top of these fundamental variables, I draw insight from option literature that shows predicting power of option pricing information for tail risks. In particular, I follow Xing et al. (2010) to construct implied volatility smirk measure, which is defined as the difference between the implied volatilities of out-of-money put option and at-the-money call option, and follow Barro and Liao (2020) to construct far-out-of-money relative option price measure, which is motivated by their pricing equation for far out-of-money put option:

$$\Omega = \frac{\alpha z_0^{\alpha} \cdot pT \cdot \epsilon^{1+\alpha-\gamma}}{(\alpha-\gamma)(1+\alpha-\gamma)} \tag{1}$$

Where Ω is the ratio of option price to stock price, and p is the probability of disaster. Since the put option price implies extreme left tail event, then it follows naturally that the counterpart measure from call option price implies extreme right tail events. This study uses both measures in the prediction model.

I use Option Metrics to construct these measures. Due to the availability of option data, I limit my sample scope between the year 1996 and 2019. Following Xing et al. (2010) and Barro and Liao (2020), I perform the following screening for put options: 1) days to expiration between 10 and 180 days; 2) implied volatility between 0.03 and 2; 3) open interest greater than zero; 4) option price greater than $0.125; 5) non-missing volume; 6) moneyness between 0.1 and 0.9. Analogously, to aid the joint prediction of jackpots, I also include the relative price of far-out-of-money call options, which obey the screening for put options, but with moneyness between 1.05 and 1.8. The option price is the mean of offer and ask prices for each option contract. The relative price of a contract is the ratio between option price and the implied forward stock price. I use open interest to calculate a weighted-average relative price. Then I average the daily relative price for each month to construct a monthly measure. I require at least 10 days of available data for each month. I use CRSP for daily and monthly stock return and volume data. Following asset pricing convention, I require common stocks in share code of 10 and 11, and with stock prices greater than $5 to avoid extreme outliers. For institutional trading, I use Thomson Reuters 13F filing data; for retail trading, I use Robintrack, which tracks Robinhood user holding of individual stocks. This dataset is available from May 2018 to August 2020. Definitions of variables are in Appendix.

3.2. Summary Statistics

The summary statistics for selected variables are shown in Table 1.

[Table 1 about here.]

As was discussed earlier, since our forecast horizon is one month, long-term historical explanatory variables might not be desirable as they may not account for regime change and hence lack sufficient flexibility (Elliott and Timmermann (2016)). Therefore, the variables are defined such that the longest lag used is one year in the case of sales growth, where I use quarter-on-quarter[6] changes to account for seasonality. All the other variables are lagged by

[6] For example, quarter one of this year on quarter one of last year.

less than 6 months, and in some cases, 3 months or even 1 month[7]. In the next section, I describe the estimation methodology for ex ante crash and jackpot probabilities, and show the baseline logit model and improved results of machine learning models.

4. Estimating Ex Ante Monthly Crash Risk

In this section, I discuss the methodologies used in both the baseline model and improved machine learning models, and show comparisons of key performance metrics for out-of-sample forecasting.

4.1. Multinomial Logit Regression

As a precursor to out-of-sample predictions, I first run an in-sample multinomial logit regression to examine whether the selected variables are strongly correlated with future realized crashes and jackpots, and whether the model is economically sound. Table 2 shows the results. Standard errors are clustered at both firm and time levels per Petersen (2009).

[Table 2 about here.]

Table 2 shows that the relative option prices of far-out-of-money puts and calls are significant predictors of crashes and jackpots in next month. Though they have the same positive sign, the coefficient on put options for crashes are greater than that for jackpots, while the coefficient on call options for crashes are less than that for jackpots. This makes intuitive sense: high relative price for far-out-of-money put options signals greater demand for protection for that particular stock, which precedes the pending crash; high relative price for far-out-of-money call options signals greater demand for speculation for that particular stock, which precedes the pending jackpot. On the other hand, surprisingly, the implied volatility SMIRK measure shows no significance in predicting crashes and jackpots. All

[7] See Appendix for variable definitions.

the other variables show coefficients in signs that are largely consistent with literature. This exercise shows that the model has substantial explanatory power with these selected variables. Next, I move on to discuss the machine learning models used to predict ex ante crash and jackpot probabilities.

4.2. Out-of-Sample Forecasting with Machine Learning

The in-sample logit shows significant explanatory power. However, it is long known that in-sample fit has substantial overfitting problem that leads to poor out-of-sample performance. Exacerbating the issue is that crashes and jackpots are rare events, and thus the estimation constitutes a "imbalanced learning" problem, where a plain logit model would produce biased estimates due to the poor finite sample properties.

To address these issues, I follow prior literature and conduct a rolling window estimation procedure, where I use 6 months of data as the training sample and 1 month data as the test sample in each window. For example, the first window consists of training sample from January 1996 to June 1996, and test sample of July 1996; the second window consists of training sample from February 1996 to July 1996, and test sample of August 1996, and so on. This procedure produces true out-of-sample estimates of crash and jackpot probabilities for next month. To improve the forecasting power, I use logistic Ridge regression as the main model. There are several reasons that Ridge is chosen: first, it is logistic regression based, and hence an easy extension from the logit model; the model produces interpretable coefficients; we are able to tune the model by the penalty factor λ to search for the best estimator[8]. The multinomial logistic Ridge seeks to estimate:

$$Pr(G = k|X = x) = \frac{\exp \beta_{0k} + \beta_k^T x}{\sum_{l=1}^{K} \exp \beta_{0l} + \beta_l^T x} \tag{2}$$

Where K is number of classes. The general elastic net (Zou and Hastie (2005)) penalized

[8] In robustness tests, I show that using other machine learning models such as LASSO, Elastic Net, and Support Vector Machines produce similar results.

negative log-likelihood function can be written as:

$$\begin{aligned}\ell(\{\beta_{0k}, \beta_k\}_1^K) = -[\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - \log(\sum_{l=1}^{K}\exp\beta_{0l} + x_i^T\beta_l))] \\ + \lambda[\frac{1}{2}(1-\alpha)\|\beta\|_F^2 + \alpha\sum_{j=1}^{p}\|\beta_j\|_q]\end{aligned} \tag{3}$$

Where λ is the penalty factor for the weighted L-1 and L-2 penalties, α is the weight of L-1 penalty. Hence Ridge regression is a special case when $\alpha = 0$. L-2 penalty is particularly suitable in our setting, since model sparsity is not a concern (number of variables are far less than number of observations).

In each rolling window, I split the training sample into two parts: first 5 months as the training set, and the last 1 month as the validation part[9]. Then I use the training set to tune the penalty factor λ of the Ridge model, and use the validation set to find the best Ridge estimator. Then this estimator is used to fit the test sample in the same window. This rolling window data split scheme can be represented in Figure 1.

[Fig. 1 about here.]

As shown in the figure, the bars represent months of data in a window. From top to bottom: the first bar represents the training set, which consists of five months of data; the second bar is the validation set, consisting one month of data; the last bar is the test set, consisting one month of data. For example, the first rolling window consists training data (January 1996 to May 1996), validation data (June 1996), and test data (July 1996). The training set is used to fit the model; the validation set is used to tune the hyper parameters to find the best estimator in terms of forecasting metric (e.g., F1 score); the resulting optimal

[9] Cross validation is usually used in training machine learning models, where the data is assumed to be i.i.d. However, in this setting, the data is in a panel structure, where observations might show substantial temporal structure. This time structure contains valuable information, and hence generic cross validation would ignore this structure and hence produce possibly inferior results. See for example Roberts, Bahn, Ciuti, Boyce, Elith, Guillera-Arroita, Hauenstein, Lahoz-Monfort, Schröder, Thuiller, Warton, Wintle, Hartig, and Dormann (2017) for detailed discussion. Nonetheless, in robustness tests, I show that by using three fold cross validation, the results are not materially altered.

estimator is then used to fit the test data, and compare the prediction with the ground truth, in order to generate the test performance metrics. As a comparison, I also apply the simple logit model on the whole training sample (6 months), and use the coefficients to directly fit the test sample.

One issue remains: since crashes and jackpots are rare events, the usually logistic estimator would produce biased estimates due to the poor finite sample properties[10]. To address this issue, I introduce a widely used machine learning technique: Synthetic Minority Oversampling Technique (SMOTE), introduced in the seminal paper by Chawla et al. (2002). They show that a combination of under-sampling majority class and over-sampling minority class can effectively improve classification performance for severely imbalanced samples. Oversampling is achieved by creating synthetic observations along the lines in the feature space that join the minority class K-nearest neighbors[11]. More formally, let $x_{minority}$ be an observation of minority class observed in the training sample, let $\widetilde{x}_{minority}$ be a random neighbor sampled adjacent to $x_{minority}$. Then a synthetic minority observation can be generated as in Equation 4:

$$x^{syn}_{minority} = w \cdot x_{minority} + (1 - w) \cdot \widetilde{x}_{minority} \tag{4}$$

Where $w \in (0, 1)$ is a random number. The $k-$nearest neighbors are sampled repeated with replacement, and corresponding synthetic observations are created until the desired balance between minority and majority classes are achieved[12].

Regardless of the intricacies of the model, the intuition is clear: assuming the features of the minority class are sufficiently clustered, it is reasonable to create "similar" observations within that cluster. In this paper, in order not to lose information of the majority class, I use SMOTE to create synthetic observations for both crashes and jackpots using oversampling, while keeping all "plain" examples without under-sampling. I show that using this technique

[10] See for example King and Zeng (2001) for discussion.

[11] See Friedman, Hastie, and Tibshirani (2001) for introduction to K-nearest neighbors.

[12] In our case, balance means that crashes, jackpots, and plain cases have the same number of (synthetic) observations.

can greatly improve the metrics for crashes and jackpots.

I follow machine learning literature and choose the following metrics: precision, recall, and F1-score. (Seliya, Khoshgoftaar, and Van Hulse (2009)). The common accuracy measure that is used in most forecasting literature is not suitable in imbalanced sample classifications (Batista, Prati, and Monard (2004)). Therefore I do not report accuracy. The definitions for precision, recall, and F1-score are as follows:

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives} \tag{5}$$

$$Recall = \frac{True\,Positives}{True\,Positives + False\,Negatives} \tag{6}$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

These metrics are computed for each of the three classes: crash, plain, and jackpot. Since I use rolling window estimations, each rolling window exercise can generate a set of metrics that evaluate out-of-sample performance, then I compute the mean metrics. There are in total 281 rolling windows and the same amount of associated sets of metrics. I summarize the mean metrics for simple logit and Ridge in Table 3.

[Table 3 about here.]

Table 3 show that across the board, especially in crash and jackpot categories, Ridge regression shows far superior performance than the simple logit. For example, the recall of crashes on average improves by a factor of nearly 70, while the F1 score of crashes on average improves by a factor of around 2.5. In the case of jackpots, the recall improves by a factor of 25, while the F1 score improves by a factor of 6. It is important to note that precision for both tail classes suffer a bit, though if looked at alone, they are misleading in that it only cares about how many observations are true out of all the predicted observations in that class. In our case, the recall measure is more important, as it identifies the model's ability to capture the true classes as much as possible. F1 score seeks to balance the two measures, and provides a more nuanced view of the model's power.

To visually demonstrate the comparison of metrics between models, I also plot the confusion matrices for the two models in the aggregate sense, where I simply add up predicted classes across time. A confusion matrix is a square matrix, where the rows are designated as true classes, and the columns are designated as predicted classes. Hence the diagonal elements are true classes that are successfully predicted. Then it follows that if we normalize the matrix row by row, the diagonal elements can be viewed as recall for each class. Figure 2 plots the matrices for all models.

[Fig. 2 about here.]

As shown in Figure 2, the Ridge model performs substantially better than the simple logit, as it is shown that The gravity of each class is more heavily concentrated along the diagonal, which is more ideal.

As an illustration of the predicted probabilities, I plot the monthly mean crash and jackpot probabilities in Figure 3.

[Fig. 3 about here.]

On top of the improved out-of-sample performance, the Ridge model seems to separate the left and right tails pretty well: the unconditional correlation between crash and jackpot probabilities is around -25%, while prior studies (for example, Conrad et al. (2014)) often show strong positive correlation between the two tails. Overall, machine learning models combined with SMOTE produce far superior out-of-sample results as compared to simple logit. Armed with a more convincing set of estimates, I turn to implications of monthly crash risk for cross-section of stock returns.

5. Are Monthly Crash Risk Priced?

In this section, I examine whether ex ante monthly crash risk is priced in the market. Literature has long shown that tail risk is priced[13], as investors have great hedging demand against extreme tail events. Prior studies such as Conrad et al. (2014) and Jang and Kang (2019) show that the two tails are likely negatively priced, as investors follow positive feedback strategies, which renders these lottery like stocks overpriced, and they subsequently experience lower returns. As they focus on the next year's crash risk, this paper studies more short-term phenomenon, where I estimated firm ex ante monthly crash risk, jointly with jackpot risk. Following the same rationale, we would expect these risks to be priced in the market. I proceed first in a portfolio test, and then examine the issue in the cross section.

5.1. Time-Series Portfolio Tests of Monthly Crash Risk

I run time-series portfolio return regressions on time-series factors benchmarking various asset pricing models. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French (1993)), then augmented with a momentum factor (FF4) (Carhart (1997)), Fama-French five-factor model (FF5) (Fama and French (2015)), and finally FF5 augmented with momentum factor (FF6). At the end of each month, I sort the stocks based on their predicted next-month crash probabilities from the Ridge model into decile portfolios, then I calculate either equal-weighted or value-weighted portfolio returns for the top decile and bottom decile, and form a zero-cost trading strategy by longing the top decile and shorting the bottom decile, and regress the excess returns on pricing factors. I apply common asset pricing filters to the stocks: stocks with a share code of 10 or 11, and month-end price of greater than \$5. The results are shown in Table 4.

[Table 4 about here.]

[13] See for example Kelly and Jiang (2014).

As shown in Table 4, when we long highest crash risk decile portfolio and short lowest decile portfolio, we produce consistent and significant negative alphas across different asset pricing models, equal-weighted or value-weighted, with t-statistics of magnitude of well over 3. Next I show more detailed results for the ten decile portfolios, both value-weighted and equal-weighted, to examine the return behavior of crash risk. The results are shown in Table 5.

[Table 5 about here.]

As shown in Table 5, no what which asset pricing model we choose, the value-weighted decile portfolio alphas largely decrease monotonically from lowest decile in crash risk to highest decile. It shows that monthly crash risk is negatively priced, consistent with the annual measures of crash risk in prior literature.

5.2. Monthly Crash Risk and Cross-Section of Stock Returns

Next, I examine the relationship between monthly crash risk and cross-section of stock returns. I run Fama-MacBeth regressions (Fama and MacBeth (1973)) following the procedure in Fama and French (2020), where I regress raw stock returns on cross-sectionally centered lagged firm characteristics. Then the coefficients on characteristics can be directly interpreted as average priced return spread of one standard deviation of the corresponding firm risk. I include common risk characteristics such as size, book-to-market ($B2M$), asset growth (ATG), profitability (ROE), momentum (MOM), short-term reversal (REV), and my estimated crash probability and jackpot probability from Ridge. On top of these variables, I also follow Jang and Kang (2019) and control for a battery of anomaly characteristics that are shown to be significantly correlated with future stock returns: abnormal capital investment ACI (Titman, Wei, and Xie (2004)), illiquidity $ILLIQ$ (Amihud (2002)), turnover $TURN$, idiosyncratic volatility $IVOL$, asset growth AG per Cooper, Gulen, and Schill (2008), composite equity issues CEI (Daniel and Titman (2006)), gross profitability

GP (Novy-Marx (2013)), net operating assets NOA (Hirshleifer, Hou, Teoh, and Zhang (2004)), net stock issues NSI (Ritter (1991)), and O-score $OSCR$ (Ohlson (1980)). Finally, there might be concern that, since the crash and jackpot probabilities are estimated from a log transform of linear combination of a series of variables, the coefficients on crash and jackpot probabilities might just represent the price that investors pay for the underlying variables. Hence in the last specification, on top of all the control charactersitics, I also include all the predictor variables that I use in the Ridge regressions. I report the regression results in Table 6.

[Table 6 about here.]

Table 6 show that even after controlling for common risk characteristics and a plethora of anomaly variables, and finally the whole set of predictor variables, the loadings on ex ante monthly crash risk remains economically and statistically significant, with comparable magnitude with the time-series portfolio alpha results. One-standard-deviation change in ex ante monthly crash risk predicts negative return spread between -0.731% to -0.258%. The coefficient on jackpot risk is largely positive, but turns negative and insignificant in the last specification, which suggests that further study needs to be done in future with respect to the right tail. Nevertheless, these results provide strong support for the efficacy of the prediction models, and consistent evidence that ex ante monthly crash risk is robustly priced in the market. Next, I turn to institutions and retail investors to explore how their trading behavior with respect to the left tail.

6. Institutional and Retail Trading on Crash Risk

Prior literature[14] show evidence that institutional investors tend to "ride the bubble" as rational speculators, instead of trading against one-year ahead crash risk as rational arbitragers. They argue that such behaviors may drive the stock prices further away from

[14] See Conrad et al. (2014) and Jang and Kang (2019).

fundamentals, exacerbating the bubble conditions à la De Long et al. (1990a), De Long et al. (1990b), and Abreu and Brunnermeier (2003). By further subsetting institutions, literature suggests that inexperienced institutional investors may ride the bubble and subsequent crash not due to rational speculation, which nevertheless help the formation and burst of the bubble (Greenwood and Nagel (2009)), showing heterogeneity among institutions with respect to their trading behavior. One relevant question is, since the estimated ex ante crash risk in these settings are one-year ahead estimates for the longer run, whether institutions have the ability to time the crash is uncertain. An immediate second question is, if institutions are sophisticated in trading crash risks, then they should be equally likely to identify jackpot stocks and earn abnormal profits, subject to the requirement that both crash risk and jackpot risk are measured reasonably well and can be timed in practice with reasonable time frame.

On the other side of the spectrum, as much been assumed that retail investors are by and large "noise traders" that could trade too much (Barber and Odean (2000)), and that those speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar (2013)), recent evidence from Robinhood traders show that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber et al. (2020)), while there is also evidence that mimicking portfolios based on the characteristics of "Robinhood traders" do not seem to underperform the market, but instead could be a market stabilizing force (Welch (2020)). These seemingly conflicting results point to the difficulties in characterizing retail investors' behaviors as a whole, and one wonders if both institutions and retail investors are at the upper hand of the bargain, who are the invisible losers? Are these so called "Robinhood traders" well equipped to discern crashes and jackpots?

In this section, I try to more systematically trace the trading behaviors of both institutions and retail investors, and explore a possibly quasi-experimental setting to infer the potential causal effect that retail investors may have the ability to destabilize or stabilize stock prices.

6.1. Institutions and Retail Traders: A Comparison

I first examine the trading behaviors of both institutions and retail investors in a panel setting with respect to common risk factors, together with crash risk and jackpot risk. I follow much of the literature to use Thomson Reuters 13F database to infer institutional trading. I use two measures of institutional trading. One measure is the overall change in the percentage of shares held by all institutions for each stock from the last quarter to the current quarter, defined as in Equation 8:

$$Inst\%Change_{i,t} = \frac{Shares\,Held\,by\,Inst_{i,t} - Shares\,Held\,by\,Inst_{i,t-1}}{Total\,Shares\,Outstanding_i} \tag{8}$$

Where t is in quarters. In the second measure, I follow Grinblatt, Titman, and Wermers (1995), Wermers (1999), and Jiang (2010), and construct an institutional trading imbalance measure as in Equation 9:

$$Inst\%Imbalance = \frac{Number\,of\,Net\,Buyers_{i,t} - Number\,of\,Net\,Sellers_{i,t}}{Total\,Number\,of\,Institutions\,holding\,the\,stock_{i,t}} \tag{9}$$

Where an institution j is a net buyer of stock i in quarter t if $Shares\,held_{i,j,t} - Shares\,held_{i,j,t-1} > 0$, and is a net seller if $Shares\,held_{i,j,t} - Shares\,held_{i,j,t-1} < 0$. This measure intuitively tracks the percentage of net institutional buyers or sellers in each quarter for a particular stock.

For retail investors, I construct retail trading imbalance measure from Robintrack data. As has been extensively discussed in Barber et al. (2020) and Welch (2020), Robintrack data contains hourly stock popularity numbers that are measure by how many users on Robinhood hold a particular stock at certain hour. Since we cannot observe the number of shares they hold for each stock, and there is no data for total number of users for each time period, the next best thing we can do is to measure the change in number of users for each stock. As my risk measures for crashes and jackpots are estimated at monthly frequency, I use month-end numbers of Robinhood users to merge the data. Therefore the measure for retail trading can be constructed as in Equation 10:

$$Change\#User = \log(\#User_{i,t}) - \log(\#User_{i,t-1}) \tag{10}$$

Where t is at monthly frequency.

Armed with these measures, I now explore their trading behaviors in a panel setting. The institutional sample runs from 1996 to 2019 at quarterly frequency, while the Robinhood sample runs from May 2018 to November 2019, subject to the data limitation. On top of ex ante monthly crash risk and jackpot risk measures, I control for a plethora of common risk characteristics as additional explanatory variables. These include: size, excess return over the market over the last quarter, detrended turnover over the last quarter, asset growth rate over the last quarter, tangible assets, sales growth, ROE of the most recent quarter, firm age, book-to-market ratio. I also add the following as additional controls: betas of Fama-French 3-factor models by running daily regressions of excess returns on these factor returns over the last quarter; idiosyncratic volatility as the residual volatility obtained from the above regressions; and total volatility of the stock over the last quarter. These measures should proxy for investors' preferences over common risk factors and also reflect possible style changes of institutions over time.

I first examine institutional trading behavior by using trading imbalance as a proxy. I first run Fama-Macbeth cross-sectional regressions and average the time-series coefficients. To control for possible unobserved heterogeneities, I also run a panel fixed effects model, with both firm and time fixed effects. I show the results in Table 7. All explanatory variables are cross-sectionally winsorized at [0.5%, 99.5%] level to remove the effects of outliers.

[Table 7 about here.]

Consistent with prior literature, Table 7 shows strong evidence that more institutional investors are net buyers of high ex ante crash risk stocks at the end of each reporting quarter, even after controlling for a plethora of firm characteristics. This trading behavior is likely to push the price of these stocks high, rendering the returns lower in the subsequent month, consistent with the negative price of the risk that I show in the last section. At the same time, more institutions seem to also be net buyers of high jackpot risk stocks, consistent with the findings in Conrad et al. (2014), though the price of ex ante jackpot risk is not conclusive in this study. For the sake of brevity, I omit the results from using holdings change measure, which are largely consistent with the findings here.

Next, I examine the trading behavior of retail investors using the imbalance measure inferred from Robintrack data. Following the prior procedure, but at a monthly frequency, I first run Fama-MacBeth regressions of retail trading imbalance on ex ante monthly crash and jackpot risks, controlling for other characteristics, and then examine the results from a panel setting, where I add firm and time fixed effects. The results are shown in Table 8.

[Table 8 about here.]

Though not significant in the Fama-MacBeth regressions, the coefficients on both crash and jackpot risks are positive and significant in the panel regression, suggesting that retail investors are chasing both tails, consistent with prior literature that they have a preference for lottery-like stocks. Jointly with the evidence shown from the trading behavior of institutional investors, these results show that investors in general buy up high ex ante monthly crash risk stocks and high jackpot risk stocks, rendering these stocks overpriced, and subsequently leading to negative return spread in the next month.

One interesting question that remains to be answered is, whether institutions are piggybacking on retail investors' trading behavior, such that they are riding the bubble, or they are themselves positive feedback traders per De Long et al. (1990b). Unfortunately due to data limitation, it remains to be studied in greater detail the timing of institutional trades and retail trades. From the results shown so far, one thing can be more or less certain: at least some of these investors are positive feedback traders, while others are possibly riding the bubble.

6.2. The Impact of Retail Trading on Crash Risk

There has been much debate in literature whether and how much retail investors can affect stock prices. Classical asset pricing models assume rational investors are price takers, and there is no room for price impact (Merton (1973)). Recent evidence suggests that retail investors do affect stock volatility (Foucault et al. (2011)); they may be marginal price setters for small stocks (Graham and Kumar (2006)); retail short sellers predict negative future returns, and they seem to have superior knowledge of small firm fundamentals (Kelley and Tetlock (2017)). Much of the literature focus on predictive tests, as it is extremely difficult to find ideal settings for proper identification for any claims for causality. Foucault et al. (2011) was one of the papers that use quasi-natural experiments to identify the causal effect of retail trading on stock volatility.

Another strand of literature that is relevant to this study is the feedback effect between option trading and stock trading, as two significant predictors of crash and jackpot risks are far-out-of-money put and call option relative prices with respect to stock forward price. Anthony (1988) was among the first to examine the sequential information flow from options to stocks. Hence the far-out-of-money options themselves are good proxies for ex ante stock crash risk. Therefore, in subsequent tests, I look at both predicted crash risk and the far-out-of-money option variables.

In this subsection, I explore one possible quasi-natural experimental setting. Robinhood introduced commission-free option trading on its platform on December 12, 2017, which would take effect in 2018[15]. Since then, Robinhood traders appear to have developed a zeal for option trading, so much that they actively discuss their Robinhood option trading positions and gains and losses on social platform, such as Reddit. After all, option trading brings the benefit of cheap leverage. In fact, around 13% of Robinhood users trade options, according to the firm disclosure[16]. This is not a small number, considering the total users

[15] Introducing Options Trading, *Robinhood Financial LLC*, `https://blog.robinhood.com/news/2017/12/12/introducing-options-trading`.

[16] See article: New Army of Individual Investors Flexes Its Muscle,

amount to 13 million in 2020[17], and hence there are at least 1.69 million users on Robinhood actively trading options. This influx of Robinhood option trader army should drive the demand for options for popular stocks, and thus affect option prices. The trading of popular stock options should in turn transmit to the elevated trading activities in the underlying stocks. This event was not caused by underlying option or stock returns and volatilities, and hence should serve as a suitable experiment.

Therefore, I hypothesize that after the introduction of option trading, those stocks whose options experienced influx of Robinhood traders should observe their crash risk increasing, compared to those similar stocks that do not have this influx around the event. And the source of the increase might come from increased demand of far out-of-money options. This increased trading of options shuold also translate into increase trading of the underlying stocks. One difficult issue, however, is that there is no direct way to identify which stocks experienced influx of Robinhood traders with respect to their options. Even though we do observe which stocks are popular among Robinhood traders, but unfortunately Robinhood do not share their option trading data.

To circumvent this issue, I explore textual information from a popular online social media platform: Reddit and its particularly popular subreddit called "WallstreetBets"[18]. As of January 2021, this subreddit has 1.8 million total active users, who post regularly everyday. I explore two "flairs" in this subreddit: "daily discussions" and "what's your move tomorrow?". I choose these two flairs because users post here every trading day, such that I have a steady number of posts and comments. I scraped all the first-level and second-level comments each day from December 2017 to September 2020. These comments are short in nature, with colorful languages. I perform two layers of pre-processing: first, I find out all the posts that contain valid ticker names. I discard those tickers that are also

Wall Street Journal, December 30, 2020, `https://www.wsj.com/articles/new-army-of-individual-investors-flexes-its-muscle-11609329600?mod=searchresults_pos2&page=1`.

[17] See `https://www.businessofapps.com/data/robinhood-statistics/`.

[18] *Wallstreetbets*: `https://www.reddit.com/r/wallstreetbets/`.

common English words, slangs, or month abbreviations (e.g. SEP). Second, I find out all the posts with tickers that mention “option”, “call’, or “put” to identify possible option buying activities. I assume that, if a user posts a comment with tickers in it, and mentions option terms in the same post, then he/she is more likely to have traded in these options, which is a reasonable assumption. Through this methodology, I can identify which stocks are likely to experience sudden influx of retail traders with respect to both options and underlying stocks.

To illustrate the extent to which they mention stocks and options in their comments, for each day in the sample, I summarize the number of unique posts that contain tickers, of which number of posts that mention options, number of unique firms mentioned, of which number of firms that mention options. I then plot the two series as in Figure 4 and Figure 5.

[Fig. 4 about here.]

[Fig. 5 about here.]

Subsequently, I use the firms that are co-mentioned with options in Wallstreetbets comments as a proxy that retail investors participate in the option trading of these stocks after Robinhood introduction of commission-free option trading in December 2017. Therefore, the sample can be divided as following: I restrict my attention to the year 2017 and 2018, with 2018 as post event period. The aforementioned firms will be the treatment group, and the rest with valid crash probability estimates as the control group. Then I conduct a standard difference-in-difference analysis similar to that in Foucault et al. (2011). I estimate the following equation as in Equation 11:

$$Crash\,Risk_{i,t+1} = \alpha + \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post + \gamma Controls_{i,t} + \epsilon_{i,t} \quad (11)$$

Where I use subscript t because I use 12 months of data for both before and after periods to improve test power. Specifically, I run two sets of tests: first, I run the diff-in-diff test with cluster robust standard errors per Petersen (2009), clustering on both firm and time level. Second, I add firm and time fixed effects, which would absorb the treatment and post dummies, leaving the interaction term intact. The dependent variable is the estimated ex ante monthly crash risk. *Treatment* is a dummy variable that equals one if both firm ticker and option are mentioned in comments in Wallstreetbets in 2018, and zero otherwise. *Post* is a dummy variable that equals one if the year is 2018, and zero otherwise. I also separately add controls to account for imperfect matching from possibly confounding factors. The results are reported in Table 9.

[Table 9 about here.]

As shown in table 9, the coefficient of interest is the interaction term, which accounts for the difference in treatment effect. The interaction term between *Treatment* and *Post* is significantly positive across all specifications, even after controlling for a battery of possible confounding firm characteristics. The estimated average effect is between around 1% to 1.6%, at less than 1% statistical significance level. This is strong evidence that retail participation tends to significantly increase stock ex ante monthly crash risk.

The next important question is whether the effect of retail participation is stronger in smaller firms. As is often shown in literature, retail investors are more likely marginal price setters for smaller stocks, while they are unlikely price setters for large stocks since their positions are much smaller than institutions. It follows naturally that in the case of ex ante monthly crash risk, retail investors should have a greater impact on smaller firms. To test this hypothesis, I subset the firms at the beginning of 2017 into two groups, one with market value greater than the cross-sectional median, the other lower than the median. In this way, I generate a dummy variable $Big = 1$ if it belongs to the larger cohort, or zero otherwise. Then I conduct a triple difference-in-difference analysis, where I interact *Treatment*, *Post*, and *Big* in the same setting as the prior tests, such that the triple interaction term can be interpreted as the incremental treatment effect on large firms. The resulting specification can be represented as follows:

$$\begin{aligned}Crash\,Risk_{i,t+1} &= \alpha + \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post \\ &+ \beta_3 Big + \beta_4 Post \times Big + \beta_5 Treated \times Big \\ &+ \beta_6 Treated \times Post \times Big + \gamma Controls_{i,t} + \epsilon_{i,t}\end{aligned} \tag{12}$$

As before, I first run panel regressions with clustered standard errors on both firm and time level, and then run another test with firm and time fixed effects. The results are shown in Table 10.

[Table 10 about here.]

Table 10 shows evidence that, consistent with the literature, retail participation has a larger impact on the ex ante monthly crash risk of smaller firms, while the impact on large firms is smaller on average. The coefficient on the interaction between *Treatment* and *Post* can be read as the effect on small firms, which is statistically significant and positive, which means that retail participation will on average increase the ex ante crash risk of smaller than median size firms by about 1.4% to 1.9%. The coefficient on the triple interaction between *Treatment*, *Post*, and *Big* is statistically significant and negative, which means that the retail impact on larger firms is smaller by about 0.6% to 1%.

The above results are done through examining all stocks that are available at the time in the sample with valid data. However, there might be legitimate concern that there is still underlying variables that correlate with being selected as treatment, that might confound the results. To alleviate that concern, I also perform propensity score matching before conducting the diff-in-diff analysis. Specifically, at the beginning of the sample (January 2017), I run a logistic regression of the dummy variable $Treatment \in 0, 1$ on the pertinent explanatory variables. These variables include: size, past three-month excess return, detrended turnover, total volatility, total skewness, asset growth, tangibility, sales growth, return on equity, firm age, book-to-market ratio, SMIRK, relative far out-of-money put option price, and relative far out-of-money call option price. Then I generate the propensity score for each stock based on the fitted values of the logistic regression. For each treatment stock, I find the five stocks that have the closest propensity scores to the treatment stock, and randomly select two of them, with replacement. In this way, I match each treatment stock with at least one control stock. Then I run the same specifications as before. The results are presented in Table 11.

[Table 11 about here.]

Table 11 shows that, consistent with prior results using full sample, with PSM matched control firms, the treatment stocks display increased ex ante crash risk by around 1% to 2.2%, depending on the specification. Moreover, there is consistent evidence that this effect differs between big and small firms: the effect on larger firms is around 0.6% to 1.7% less than the small firms, supporting the notion that retail investors might be marginal price setters for small firms.

Finally, it would also be interesting to see whether retail participation will impact the underlying variables that I use to predict ex ante monthly crash probabilities. This would point to some channels that could also partially drive the increase of crash risk[19]. I choose the following dependent variables to examine: the relative far out-of-money put and call option prices; trading volume as volume scaled by shares outstanding; total return volatility; and total return skewness. I follow the last test to run a triple difference-in-difference specification, with firm characteristics as controls. Therefore, the variables of interest are the interaction between $Treatment$ and $Post$, and the triple interaction between $Treatment$, $Post$, and Big. I present the results by using firm and time clustered standard errors adjustment in Table 12[20].

[Table 12 about here.]

[19] Note that in Table 11, I add predictor variables in the last two specifications as controls, and the results are still robust. But nonetheless it is interesting to examine the additional effects of retail participation on firm characteristics.

[20] I also ran panel regressions with firm and time fixed effects, and the results are largely similar. I omit them for brevity.

Consistent with intuition, across the board, there is positive treatment effect for the five variables, as shown by the interaction term between $Treatment$ and $Post$, though the coefficients for trade volume and total return skewness are not significant. In addition, all the triple interactions between $Treatment$, $Post$, and Big are shown as negative, supporting the prior finding that retail investors have a much less impact on bigger firms. One interesting results is that the relative prices of both far out-of-money put and call options are significantly increased for small firms, suggesting a larger demand for these options, but the effect for large firms is muted since the triple interaction offsets it almost entirely. There is further anecdotal evidence that on Wallstreetbets, traders often boast how they trade options on small stocks. Another interesting results is that retail participation tends to significantly increase firm's stock return volatility, consistent with the findings in Foucault et al. (2011), and the effect is smaller for larger firms.

Taken together, these results enrich our understanding of how retail investors shape the tail risks of firms. Overall, the experiment provides support that retail participation would significantly increase firm ex ante monthly crash risk, and a host of related underlying characteristics. Moreover, these effects are stronger in smaller firms, and weaker in large firms. In other words, retail investors tend to make the left tail fatter, while chasing the left tail.

7. Robustness Tests

7.1. Results Using Other Machine Learning Models

I include high-minus-low portfolio return regression results here for a set of machine learning models, where I use the same rolling window estimation procedure, but use three-fold cross validation to tune the model. The zero-cost portfolio alphas are presented in Table 13.

[Table 13 about here.]

The results show that the estimates of ex ante monthly crash risk are robust to different underlying estimating models.

8. Conclusion

This study builds on prior literature to develop an ex ante measure for firm-level monthly crash and jackpot probabilities through machine learning. I use imbalanced learning techniques with Ridge regression to show strong forecasting power for subsequent one month crashes and jackpots. The estimated crash risk is robustly priced both in time-series portfolio tests and cross-sectional tests. Armed with the estimated crash risk, I show that institutions and retail investors seem to chase the left tail, rendering the high risk stocks overpriced, and revealing a negative return spread subsequently. Using Robinhood introduction of commission-free option trading at the end of 2017 as a quasi-experiment setting, together with textual information from Reddit, I show that retail participation significantly increase ex ante stock crash risk, and this effect is stronger for small firms.

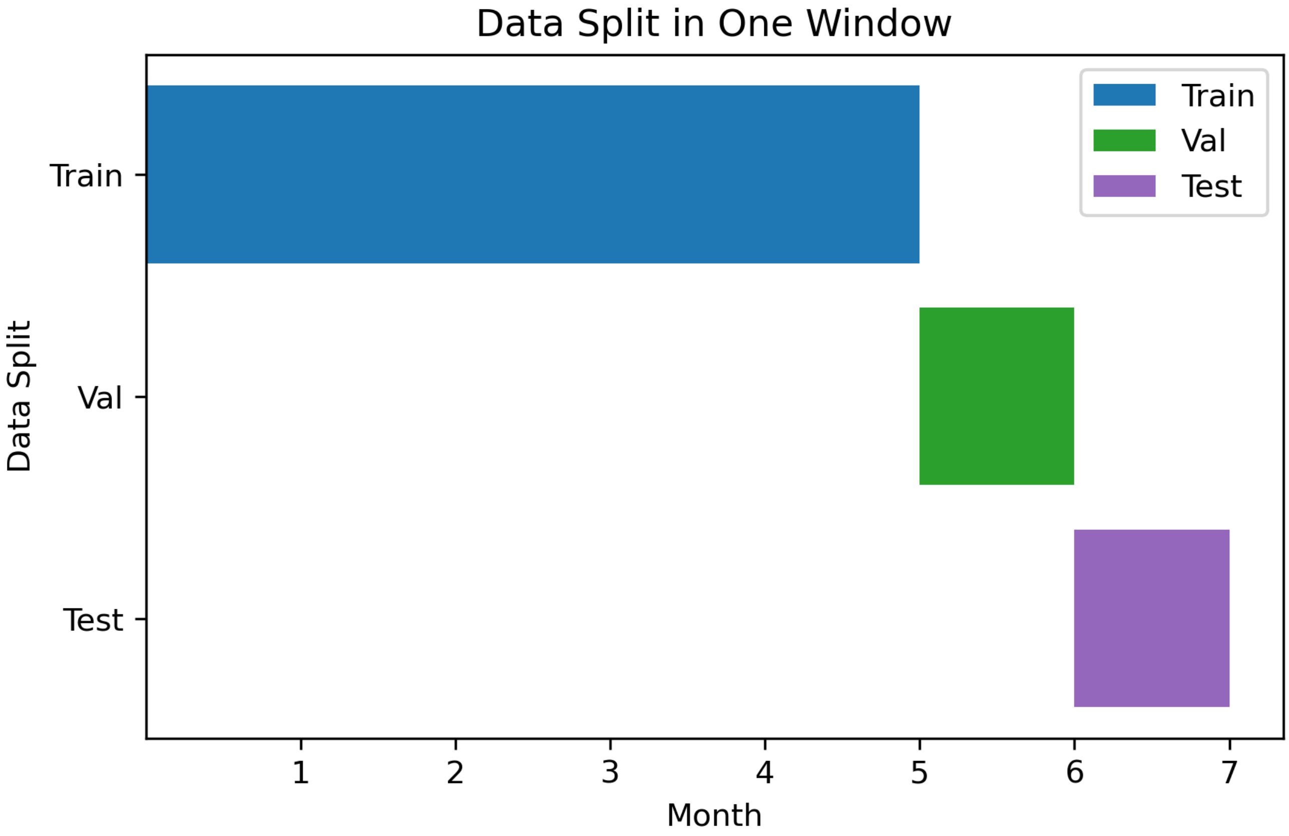

Fig. 1. Data Split in One Window. This figure plots how the data of one rolling window is split. The bars represent months of data in a window. From top to bottom: the first bar represents the training set, which consists of five months of data; the second bar is the validation set, consisting one month of data; the last bar is the test set, consisting one month of data. For example, the first rolling window consists training data (January 1996 to May 1996), validation data (June 1996), and test data (July 1996). The training set is used to fit the model; the validation set is used to tune the hyper parameters to find the best estimator in terms of forecasting metric (e.g., F1 score); the resulting optimal estimator is then used to fit the test data, and compare the prediction with the ground truth, in order to generate the test performance metrics.

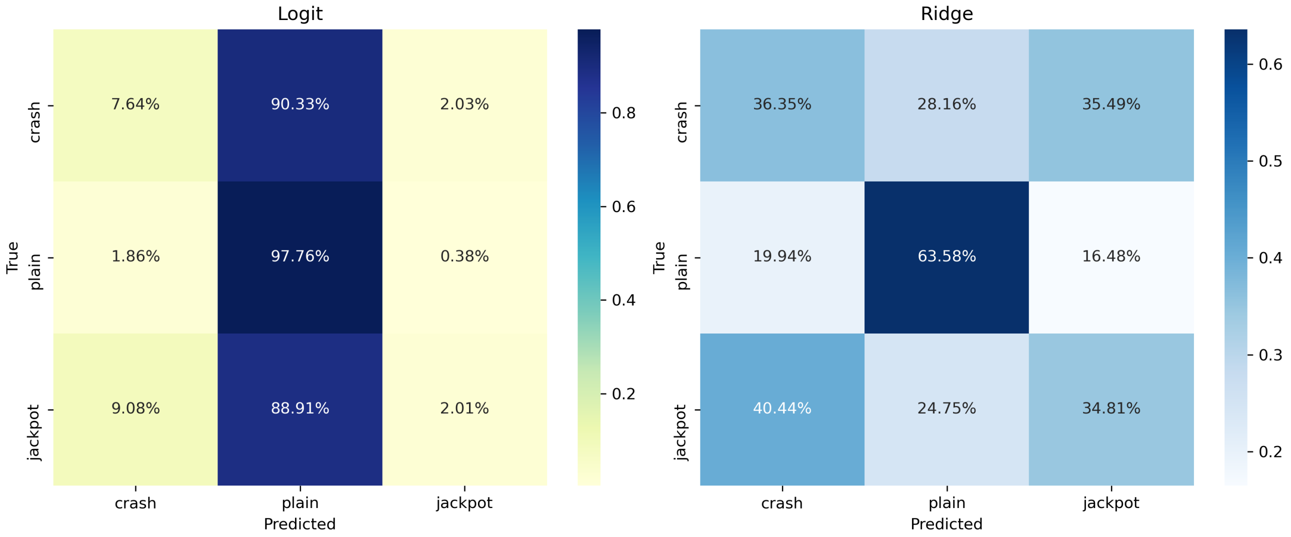

Fig. 2. Aggregate Confusion Matrices. This figure plots the aggregate confusion matrices for simple logit and Ridge, where the predicted classes are add up across time. The rows are true classes, while the columns are predicted classes. All elements are normalized row by row, such that the diagonal elements can be viewed as recall for each class.

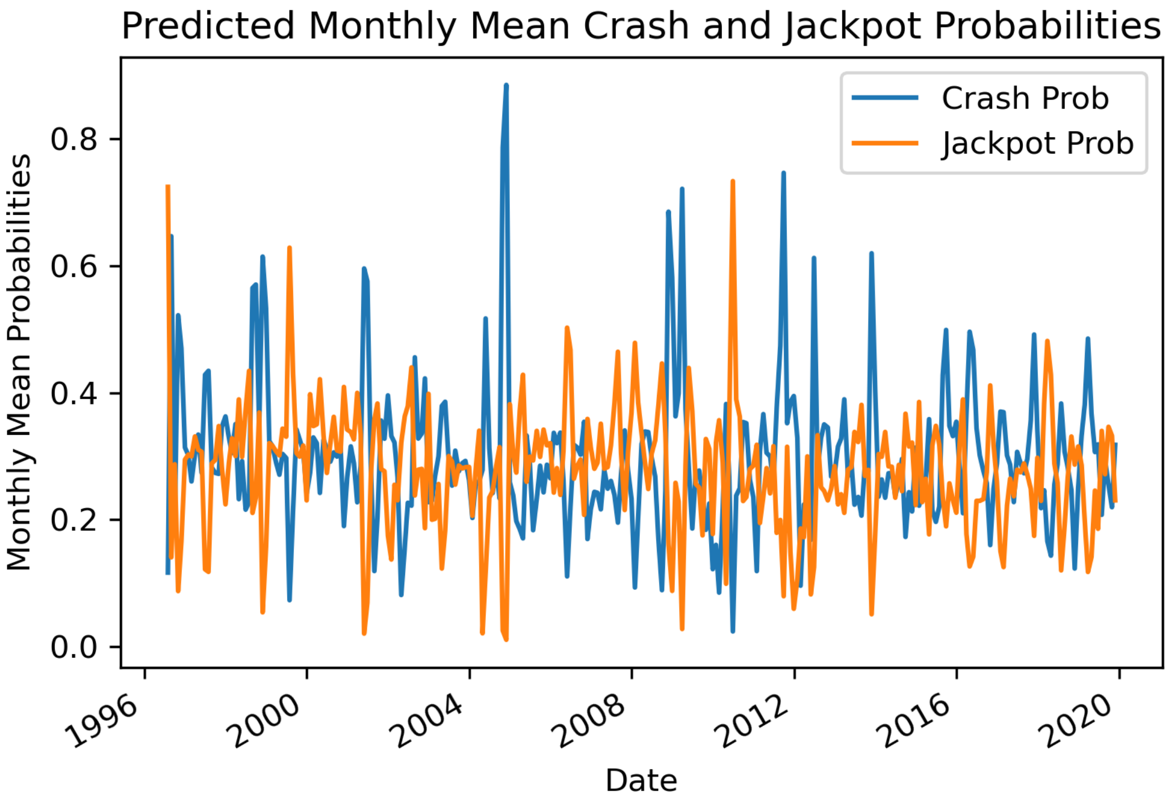

Fig. 3. Mean Monthly Predicted Crash and Jackpot Probabilities. This figure plots the mean monthly predicted crash and jackpot probabilities over time, per Ridge model. Each month, I calculate the cross-sectional mean predicted crash and jackpot probabilities respectively, and then plot them against time. The sample runs from July 1996 to December 2019.

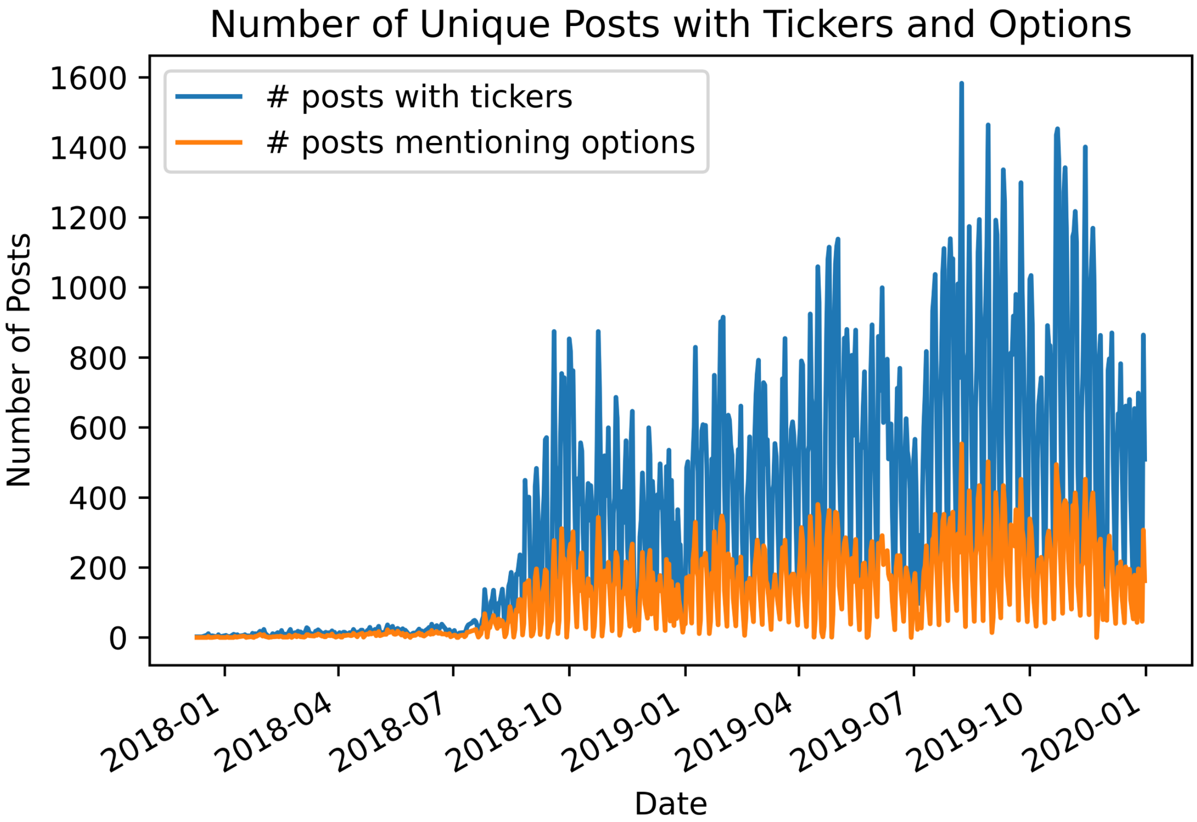

Fig. 4. Number of Posts Over Time. This figure plots the number of unique posts that contain ticker names, and of which, number of posts that mention options on Wallstreetbets of Reddit. The sample runs from December 2017 to December 2019.

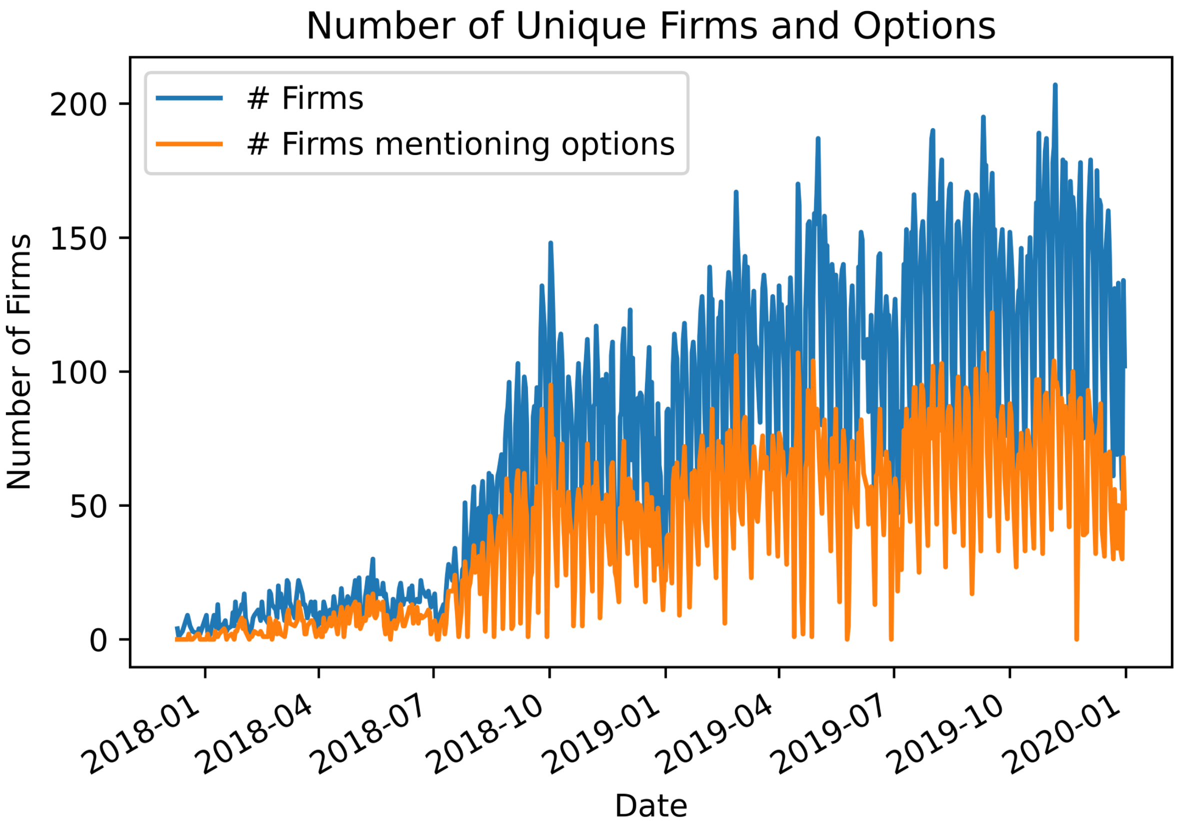

Fig. 5. Number of Firms Mentioned Over Time. This figure plots the number of unique firms that were mentioned, and of which, number of firms that are also co-mentioned with options on Wallstreetbets of Reddit. The sample runs from December 2017 to December 2019.

Table 1: Summary Statistics

The table presents summary statistics for key variables used in my baseline model for predicting crash probabilities. For sake of brevity, I present characteristics for crashes and jackpots, omitting “plain” cases. The total observations amount to 403,379 firm×month pairs, with “plain” cases at 358,780 observations. I define crashes as monthly log return of less than -20%, and jackpots as monthly log return of greater than 20%. SMIRK is the implied volatility smirk measure per Xing et al. (2010). FOMP and FOMC are far-out-of-money put option and call option relative price measure per Barro and Liao (2020). Variable definitions can be found in Appendix. The sample runs from January 1996 to December 2019. All variables are lagged properly.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	crashes				Plain				jackpots			
VARIABLES	mean	sd	min	max	mean	sd	min	max	mean	sd	min	max
Size	20.74	1.340	16.73	27.60	21.64	1.570	16.45	27.80	20.60	1.281	16.66	27.06
Exret3	-0.033	0.322	-2.295	2.436	0.007	0.193	-2.235	2.449	-0.038	0.340	-2.142	2.377
RM3	-0.004	0.096	-0.396	0.253	0.025	0.076	-0.396	0.253	0.004	0.100	-0.396	0.253
Dturn	-0.002	0.300	-10.42	10.16	0.0002	0.145	-7.975	11.31	-0.007	0.265	-6.073	13.39
Tvol	0.042	0.025	0.007	0.411	0.025	0.016	0.001	0.440	0.041	0.023	0.004	0.333
Tskew	0.081	0.935	-4.644	4.626	0.072	1.020	-4.673	4.686	0.106	0.945	-4.542	4.574
ATG	0.043	0.201	-2.968	3.029	0.0306	0.126	-2.968	3.723	0.0403	0.183	-1.817	3.723
Tang	0.315	0.422	0	5.869	0.351	0.428	0	7.933	0.322	0.413	0	6.268
Salesg	0.167	0.650	-7.857	9.270	0.110	0.403	-9.901	12.43	0.172	0.605	-9.901	7.496
ROE	-0.067	44.02	-5,958	3,162	0.026	16.35	-5,958	3,162	0.049	5.554	-49.72	534.3
Age	13.81	14.42	0	93	21.83	19.48	0	93	13.85	14.07	0	93
B2M	0.540	0.649	0	18.57	0.476	0.444	0	25.85	0.588	0.780	0.0002	32.77
SMIRK	0.068	0.078	-1.056	1.355	0.053	0.050	-1.093	1.433	0.065	0.069	-0.558	1.069
FOMP_price	0.040	0.025	0.002	0.248	0.022	0.017	0.001	0.239	0.041	0.025	0.002	0.253
FOMP_price	0.047	0.026	0.002	0.265	0.028	0.019	0.001	0.339	0.048	0.026	0.003	0.298
Obs	25,355				358,780				19,244			

Table 2: Mutlinomial Logit

The table runs a multinomial logit regression predicting crashes and jackpots for sample period 1996 - 2019. "plain" cases are set as base and are omitted. Variable definitions are shown in Appendix. Each variable is properly lagged. The crashes and jackpots are classified as one-month ahead monthly log returns of less than -20% and greater than 20% respectively. SMIRK is the implied volatility smirk measure per Xing et al. (2010). FOMP and FOMC are far-out-of-money put option and call option relative price measure per Barro and Liao (2020). Standard errors are in parentheses and are clustered at stock and month levels per Petersen (2009) and are included in parentheses.

	(1)	(2)
	Crash	Jackpot
Relative_FOMP_price	7.884***	6.380***
	(1.148)	(1.431)
Relative_FOMC_price	13.06***	17.67***
	(1.529)	(1.497)
SMIRK	-0.158	-0.640
	(0.370)	(0.392)
RM3	-2.148**	-1.386*
	(0.866)	(0.793)
Exret3	-0.0874	-0.214*
	(0.0896)	(0.114)
B2M	-0.0573	0.0488
	(0.0409)	(0.0342)
ATG	0.294***	0.203*
	(0.0856)	(0.105)
ROE	-0.524***	0.278***
	(0.0904)	(0.100)
Tvol	22.54***	19.68***
	(2.275)	(2.075)
Tskew	-0.00588	0.0111
	(0.0107)	(0.0128)
Size	-0.0957***	-0.174***
	(0.0221)	(0.0207)
Dturn	-1.129***	-1.082***
	(0.191)	(0.157)
Age	-0.0112***	-0.00860***
	(0.00143)	(0.00147)
Tang	0.0670	0.0931**
	(0.0545)	(0.0457)
Salesg	0.104***	0.156***
	(0.0365)	(0.0389)
Observations	403,379	
Pseudo R2	0.123	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Mean Performance Metrics

The table reports mean performance metrics for simple logit and Ridge across the rolling prediction windows from January 1996 to December 2019. Each window consists of 6-month training set and 1-month test set. In the case of simple logit, the whole training set is fitted and used to fit the test set. In the case of Ridge, the training set is further split into 5 months of training data and 1 month of validation data, where the training data is used to tune the Ridge estimator (through penalty factor λ), and then the best estimator is chosen to fit the test set. The metrics are defined as follows:

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives}$$

$$Recall = \frac{True\,Positives}{True\,Positives + False\,Negatives}$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$G - Mean = 2 \times \sqrt{\frac{True\,Positives}{True\,Positives + False\,Negatives} \times \frac{True\,Negatives}{True\,Negatives + False\,Positives}}$$

These metrics are computed for each of the three classes. There are in total 281 windows, and hence 281 sets of metrics are generated in total for each underlying model. These metrics are then averaged across time.

Class	Metrics	logit	Ridge
Crash	Precision	0.177	0.128
	Recall	0.062	0.412
	F1	0.049	0.128
Plain	Precision	0.891	0.935
	Recall	0.970	0.626
	F1	0.922	0.730
Jackpot	Precision	0.100	0.090
	Recall	0.014	0.344
	F1	0.018	0.108

Table 4: Decile High-Minus-Low Alphas

This table presents the high-minus-low long-short zero-cost strategy alphas, per asset pricing model, for both equal-weighted and value-weighted portfolios. At the end of each month, stocks are ranked by their ex-ante crash probabilities produced by Ridge model into ten decile portfolios each month. Then the high-minus-low return series for both equal-weighted and value-weighted returns where we long highest decile portfolio and short lowest decile portfolio, are regressed on various risk factor return series. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French (1993)), then augmented with a momentum factor (FF4) (Carhart (1997)), Fama-French five-factor model (FF5) (Fama and French (2015)), and finally FF5 augmented with momentum factor (FF6). *t*-statistics are included. Time-series regressions are estimated with Newey-West standard errors with 12 lags.

	Value-weighted		Equal-weighted	
Pricing_model	Alpha	*t*-stat	Alpha	*t*-stat
CAPM	-1.523***	-2.999	-1.594***	-3.303
FF3	-1.467***	-3.810	-1.557***	-4.259
FF4	-1.054***	-2.861	-1.125***	-3.272
FF5	-0.932***	-2.881	-1.186***	-3.567
FF6	-0.664**	-2.384	-0.898***	-3.289

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Value-Weighted Decile Alphas
Stocks are ranked by their ex-ante crash probabilities each month into ten deciles. This table presents the alphas from regressing the excess returns of the ten value-weighted decile portfolios, per each asset pricing model. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French (1993)), then augmented with a momentum factor (FF4) (Carhart (1997)), Fama-French five-factor model (FF5) (Fama and French (2015)), and finally FF5 augmented with momentum factor (FF6). Newey-West standard errors with 12 lags are included in parentheses.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
CAPM	0.142*	0.215**	0.157	0.163	-0.179	-0.310**	-0.454**	-0.652**	-0.659*	-1.381***
	(0.083)	(0.086)	(0.115)	(0.129)	(0.111)	(0.135)	(0.180)	(0.269)	(0.377)	(0.440)
FF3	0.136**	0.193**	0.124	0.151	-0.193**	-0.315**	-0.445***	-0.621***	-0.653**	-1.331***
	(0.062)	(0.076)	(0.097)	(0.119)	(0.090)	(0.124)	(0.142)	(0.203)	(0.311)	(0.344)
FF4	0.049	0.151*	0.078	0.181	-0.158*	-0.223*	-0.278**	-0.453**	-0.361	-1.005***
	(0.063)	(0.089)	(0.103)	(0.121)	(0.084)	(0.117)	(0.120)	(0.200)	(0.293)	(0.336)
FF5	-0.011	0.062	0.057	0.096	-0.173**	-0.298**	-0.265	-0.427***	-0.346	-0.944***
	(0.068)	(0.083)	(0.103)	(0.117)	(0.087)	(0.132)	(0.160)	(0.157)	(0.283)	(0.300)
FF6	-0.066	0.038	0.028	0.120	-0.150*	-0.234*	-0.156	-0.317**	-0.152	-0.730***
	(0.066)	(0.092)	(0.105)	(0.120)	(0.084)	(0.125)	(0.111)	(0.158)	(0.234)	(0.275)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: FMB Cross-Sectional Regressions

This table reports Fama-MacBeth regressions of raw returns on lagged firm characteristics in the spirit of Fama and French (2020). Independent variables are centered cross-sectionally each month. Control variables include : size, book-to-market ratio, asset growth, ROE, momentum, short-term reversal. In Column (3) and (4), I add anomaly variables: abnormal capital investment ACI (Titman et al. (2004)), illiquidity $ILLIQ$ (Amihud (2002)), turnover $TURN$, idiosyncratic volatility $IVOL$, asset growth AG per Cooper et al. (2008), composite equity issues CEI (Daniel and Titman (2006)), gross profitability GP (Novy-Marx (2013)), net operating assets NOA (Hirshleifer et al. (2004)), net stock issues NSI (Ritter (1991)), and O-score $OSCR$ (Ohlson (1980)). And in Column (5), I add all the predictor variables that I use in estimating crash and jackpot probabilities from the Ridge model. Standard errors are adjusted according to Newey-West procedures.

	(1)	(2)	(3)	(4)	(5)
	Dep Var: Returns				
Crash_prob	-0.274***	-0.258***	-0.278***	-0.332***	-0.731***
	(0.065)	(0.084)	(0.092)	(0.104)	(0.255)
Jackpot_prob	0.352***	0.369**	0.379*	0.409*	-0.622
	(0.097)	(0.172)	(0.205)	(0.232)	(0.401)
Size		-0.151	-0.068	-0.092	-0.482**
		(0.103)	(0.125)	(0.127)	(0.192)
B2M		-0.107**	-0.082*	0.023	-0.097
		(0.048)	(0.044)	(0.078)	(0.100)
ROE		0.315***	0.319***	0.354***	0.296***
		(0.036)	(0.037)	(0.044)	(0.035)
ATG		-0.033	-0.029	-0.067	-0.054
		(0.074)	(0.073)	(0.046)	(0.085)
REV		-0.156***	-0.174***	-0.152**	-0.378***
		(0.048)	(0.049)	(0.069)	(0.135)
MOM		0.067	0.074*	0.053	0.013
		(0.044)	(0.044)	(0.044)	(0.063)
Illiq			2.362	2.511	0.690
			(2.133)	(2.214)	(1.358)
Turn			-0.018	-0.055	-0.042
			(0.060)	(0.086)	(0.063)
Ivol			-0.055	-0.115	-0.462**
			(0.066)	(0.088)	(0.201)
Anomalies	NO	NO	NO	YES	YES
Predictors	NO	NO	NO	NO	YES
Observations	398,604	398,604	398,604	398,604	398,604
Avg R2	0.010	0.031	0.046	0.066	0.094
Number of groups	281	281	281	281	281
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 7: Institutional Trading Imbalance and Monthly Crash Risk

This table shows the results that examine the relationship between insitutional trading imbalance and monthly crash risk. In Column (1) to (3), I run Fama-MacBeth cross-sectional regressions to estimate the average coefficients on crash and jackpot risks, controlling for other firm characteristics. In Column (4), I run panel regression, with both firm and time fixed effects to control for unobserved heterogeneities. Institutional trading imbalance is defined as:

$$Inst\%Imbalance_{i,t} = \frac{Number\,of\,Net\,Buyers_{i,t} - Number\,of\,Net\,Sellers_{i,t}}{Total\,Number\,of\,Institutions\,holding\,the\,stock_{i,t}}$$

All variables are at [0.5%, 99.5%] level to remove the effects of outliers. The sample runs from July 1996 to December 2019 at quarterly frequency.

	(1)	(2)	(3)	(4)
	Dep Var: Inst%Imbalance			
VARIABLES	FMB			Panel
Crash_prob	0.195***	0.161***	0.240	0.167***
	(0.052)	(0.043)	(0.153)	(0.007)
Jackpot_prob	0.073*	0.063***	0.020	0.143***
	(0.038)	(0.018)	(0.041)	(0.008)
Size		-0.003**	-0.006***	0.005***
		(0.001)	(0.002)	(0.001)
B2M		-0.009***	-0.043	-0.016***
		(0.002)	(0.037)	(0.002)
ROE		-0.034***	-0.032***	-0.003
		(0.006)	(0.010)	(0.004)
ATG		0.095***	0.095***	0.051***
		(0.007)	(0.017)	(0.004)
Exret3		0.007	0.030	0.026***
		(0.004)	(0.022)	(0.002)
Ivol			-0.026	-0.005***
			(0.033)	(0.002)
Tvol			0.035	0.007***
			(0.036)	(0.001)
Dturn			0.075	0.011***
			(0.065)	(0.004)
Tang			-0.004	-0.006**
			(0.009)	(0.002)
Salesg			0.020**	0.019***
			(0.009)	(0.001)
FF3 βs	NO	NO	YES	YES
Observations	113,158	113,158	113,158	112,761
R-squared	0.042	0.066	0.099	0.232
Firm FE	NO	NO	NO	YES
Time FE	NO	NO	NO	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Retail Trading Imbalance and Monthly Crash Risk

This table shows the results that examine the relationship between retail trading imbalance and monthly crash risk. In Column (1) to (3), I run Fama-MacBeth cross-sectional regressions to estimate the average coefficients on crash and jackpot risks, controlling for other firm characteristics. In Column (4), I run panel regression, with both firm and time fixed effects to control for unobserved heterogeneities. Retail trading imbalance is defined as:

$$Change\#User = \log(\#User_{i,t}) - \log(\#User_{i,t-1})$$

The user data is from Robintrack, which provides hourly data on the number of users that hold a particular stock. All variables are at [0.5%, 99.5%] level to remove the effects of outliers. The sample runs from June 2018 to December 2019 at monthly frequency.

	(1)	(2)	(3)	(4)
	Dep Var: Retail%Imbalance			
VARIABLES	FMB			Panel
Crash_prob	-0.046	0.062	0.115	0.114***
	(0.062)	(0.052)	(0.069)	(0.027)
Jackpot_prob	0.149*	0.261**	0.403***	0.229***
	(0.075)	(0.107)	(0.136)	(0.025)
Size		0.012***	0.013***	0.020***
		(0.003)	(0.003)	(0.007)
B2M		0.005	0.002	0.013
		(0.004)	(0.004)	(0.008)
ROE		0.006	-0.010	0.012
		(0.008)	(0.006)	(0.011)
ATG		0.030***	0.029**	0.011
		(0.010)	(0.010)	(0.009)
Exret3		0.029***	0.023**	0.023***
		(0.008)	(0.009)	(0.007)
Ivol			-0.011	-0.008**
			(0.007)	(0.003)
Tvol			-0.014*	-0.021***
			(0.008)	(0.003)
Dturn			-0.001	0.005
			(0.011)	(0.011)
Tang			0.017**	-0.003
			(0.007)	(0.007)
Salesg			-0.002	-0.009**
			(0.003)	(0.005)
FF3 βs	NO	NO	YES	YES
Observations	27,159	27,159	27,159	27,105
R-squared	0.026	0.043	0.087	0.130
Firm FE	NO	NO	NO	YES
Time FE	NO	NO	NO	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: The Impact of Retail Participation on Short-Term Crash Risk

This table reports the result of a difference-in-difference analysis for the impact of retail participation on ex ante monthly firm-level crash risk. The dependent variable is the estimated ex ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. In Column (1) to (4), Standard errors are clustered at both firm and month levels. In Column (2) and (3), I add a plethora of firm characteristics; in Column (4), I add predictor variables used in estimating ex ante monthly crash risk. Column (5) adds firm and time fixed effects. Sample runs from January 2017 to December 2018. The base specification (without fixed effects) is:

$$Crash\,Risk_{i,t+1} = \alpha + \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post + \gamma Controls_{i,t} + \epsilon_{i,t}$$

	(1)	(2)	(3)	(4)	(5)
	Dep Var: Ex Ante Monthly Crash Risk				
	Clustered				FE
1.treatment	-0.032***	0.011***	0.011***	-0.001	
	(0.005)	(0.003)	(0.003)	(0.002)	
1.post	-0.087***	-0.085***	-0.085***	-0.088***	
	(0.028)	(0.029)	(0.029)	(0.029)	
1.treatment#1.post	0.016***	0.015***	0.015***	0.010***	0.009***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)
Size		-0.043***	-0.043***	-0.023***	-0.028***
		(0.003)	(0.003)	(0.002)	(0.002)
B2M		-0.001	-0.002	-0.008	-0.003
		(0.006)	(0.006)	(0.005)	(0.004)
ROE			-0.000***	-0.000***	-0.000*
			(0.000)	(0.000)	(0.000)
ATG			-0.004	-0.007	-0.008**
			(0.010)	(0.010)	(0.004)
Exret3			-0.016	-0.019	-0.027***
			(0.019)	(0.018)	(0.003)
Predictors	NO	NO	NO	YES	YES
Observations	39,482	39,482	39,482	39,482	39,411
R-squared	0.120	0.416	0.416	0.532	0.868
Firm Cluster	YES	YES	YES	YES	NO
Time Cluster	YES	YES	YES	YES	NO
Firm FE	NO	NO	NO	NO	YES
Time FE	NO	NO	NO	NO	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10: The Impact of Retail Participation on Crash Risk: Big vs Small Firms
This table reports the result of a triple difference-in-difference analysis for the impact of retail participation on ex ante monthly firm-level crash risk for big and small firm cohorts. The dependent variable is the estimated ex ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. In Column (1) to (3), Standard errors are clustered at both firm and month levels. In Column (2) and (3), I add a plethora of firm characteristics. Column (4) adds firm and time fixed effects. Sample runs from January 2017 to December 2018.

	(1)	(2)	(3)	(4)
	Dep Var: Ex Ante Monthly Crash Risk			
	Clustered			FE
1.treatment	0.006	0.011***	0.012***	
	(0.005)	(0.004)	(0.004)	
1.post	-0.096***	-0.090**	-0.090**	
	(0.032)	(0.033)	(0.033)	
1.treatment#1.post	0.014***	0.019***	0.018***	0.015***
	(0.004)	(0.004)	(0.005)	(0.002)
1.big	-0.108***	-0.013*	-0.014*	
	(0.007)	(0.007)	(0.007)	
1.treatment#1.big	-0.023***	0.001	0.001	
	(0.007)	(0.005)	(0.005)	
1.post#1.big	0.016	0.012	0.012	0.012***
	(0.014)	(0.013)	(0.014)	(0.001)
1.treatment#1.post#1.big	-0.006	-0.010**	-0.010**	-0.006**
	(0.004)	(0.004)	(0.004)	(0.003)
Size		-0.041***	-0.041***	-0.051***
		(0.003)	(0.003)	(0.003)
B2M		-0.001	-0.002	0.007
		(0.006)	(0.006)	(0.006)
ROE			-0.000***	-0.000*
			(0.000)	(0.000)
ATG			-0.005	-0.007*
			(0.010)	(0.004)
Exret3			-0.017	-0.026***
			(0.019)	(0.003)
Observations	39,482	39,482	39,482	39,411
R-squared	0.298	0.416	0.417	0.840
Firm Cluster	YES	YES	YES	NO
Time Cluster	YES	YES	YES	NO
Firm FE	NO	NO	NO	YES
Time FE	NO	NO	NO	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: The Impact of Retail Participation on Crash Risk: PSM Approach
This table reports the result of various difference-in-difference analyses for the impact of retail participation on ex ante monthly firm-level crash risk for big and small firm cohorts, by using propensity score matching. The dependent variable is the estimated ex ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. Each treatment stock is matched with at least one control firm, based on propensity score matching. The propensity scores are generated by logistic regression of treatment dummy on firm characteristics at the beginning of the sample. In Column (4) and (5), control variables are added. Sample runs from January 2017 to December 2018.

	(1)	(2)	(3)	(4)	(5)
	Dep Var: Ex Ante Monthly Crash Risk				
	PSM matched				
1.treatment	-0.004	0.010		0.014**	
	(0.005)	(0.007)		(0.006)	
1.post	-0.087***	-0.101***		-0.095***	
	(0.026)	(0.031)		(0.031)	
1.treatment#1.post	0.017***	0.018***	0.010***	0.022***	0.015***
	(0.003)	(0.006)	(0.003)	(0.006)	(0.003)
1.big		-0.109***		-0.018**	
		(0.009)		(0.008)	
1.treatment#1.big		-0.022**		-0.008	
		(0.009)		(0.006)	
1.post#1.big		0.023*	0.018***	0.019	0.015***
		(0.013)	(0.002)	(0.013)	(0.002)
1.treatment#1.post#1.big		-0.014**	-0.006	-0.017**	-0.009***
		(0.006)	(0.003)	(0.006)	(0.003)
Controls	NO	NO	NO	YES	YES
Observations	19,584	19,584	19,574	19,584	19,574
R-squared	0.111	0.322	0.832	0.449	0.844
Firm Cluster	YES	YES	NO	YES	NO
Time Cluster	YES	YES	NO	YES	NO
Firm FE	NO	NO	YES	NO	YES
Time FE	NO	NO	YES	NO	YES
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 12: The Impact of Retail Participation on Crash Related Characteristics
This table reports the result of a triple difference-in-difference analysis for the impact of retail participation on ex ante monthly characteristics for big and small firm cohorts. The dependent variables include: the relative far out-of-money put and call option prices; trading volume as volume scaled by shares outstanding; total return volatility; and total return skewness. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. Standard errors are clustered at both firm and month levels. Sample runs from January 2017 to December 2018.

	(1)	(2)	(3) Dep Vars:	(4)	(5)
VARIABLES	FOMP	FOMC	Trade_Vol	Tvol	Tskew
1.treatment	0.004***	0.005***	1.327***	0.003***	0.025
	(0.001)	(0.001)	(0.256)	(0.001)	(0.017)
1.post	0.001***	0.000	0.041	0.003**	-0.037
	(0.000)	(0.001)	(0.088)	(0.001)	(0.042)
1.treatment#1.post	0.003***	0.003***	0.332	0.003***	0.003
	(0.001)	(0.001)	(0.238)	(0.001)	(0.027)
1.big	0.002**	0.001	0.588***	-0.000	-0.011
	(0.001)	(0.001)	(0.125)	(0.001)	(0.041)
1.treatment#1.big	-0.000	-0.001	-0.865***	-0.001	-0.021
	(0.001)	(0.002)	(0.274)	(0.001)	(0.035)
1.post#1.big	-0.000	-0.000	0.061	0.001	-0.011
	(0.000)	(0.000)	(0.058)	(0.001)	(0.058)
1.treatment#1.post#1.big	-0.002**	-0.003**	-0.230	-0.002**	-0.017
	(0.001)	(0.001)	(0.226)	(0.001)	(0.041)
Controls	YES	YES	YES	YES	YES
Observations	39,482	39,482	39,482	39,482	39,482
R-squared	0.350	0.339	0.062	0.213	0.078
Firm Cluster	YES	YES	YES	YES	YES
Time Cluster	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Decile High-Minus-Low Alphas
Stocks are ranked by their ex-ante crash probabilities each month into ten deciles. This table presents the high-minus-low long-short zero-cost strategy alphas, per model, for both equal-weighted and value-weighted portfolios. t-statistics are in the even rows. Time-series regressions are estimated with Newey-West standard errors with 12 lags.

model	Logit	LASSO	Ridge	Elastic Net	Linear SVM	Logit	LASSO	Ridge	Elastic Net	Linear SVM
	Equal-Weighted					Value-Weighted				
CAPM	-1.587	-1.532	-1.379	-1.519	-1.695	-1.665	-1.627	-1.488	-1.473	-1.626
t	-3.144	-2.842	-2.517	-2.995	-3.528	-3.127	-3.027	-2.809	-2.802	-3.068
FF3	-1.526	-1.476	-1.311	-1.466	-1.636	-1.611	-1.578	-1.434	-1.433	-1.565
t	-4.737	-3.678	-3.328	-3.921	-5.109	-4.421	-3.757	-3.481	-3.471	-4.208
FF4	-1.149	-1.124	-0.965	-1.105	-1.268	-1.200	-1.199	-1.062	-1.029	-1.172
t	-3.692	-3.102	-2.697	-3.418	-3.972	-3.316	-2.926	-2.690	-2.709	-3.151
FF5	-0.770	-1.013	-0.823	-0.989	-0.936	-0.770	-0.994	-0.837	-0.792	-0.764
t	-2.718	-2.615	-2.289	-2.591	-3.505	-2.452	-2.710	-2.383	-2.012	-2.392
FF6	-0.537	-0.786	-0.600	-0.756	-0.706	-0.517	-0.752	-0.602	-0.536	-0.522
t	-2.289	-2.442	-1.970	-2.531	-3.090	-1.867	-2.280	-1.933	-1.623	-1.776

Appendix A. Replications

A.1. Replicating Jang and Kang (2019)

I replicate the main results of Jang and Kang (2019) for the sample period 1996 - 2019. First, I confirm the main results of multinomial logistic regression of exploring the relationship between crashes and jackpots and various firm characteristics. Table A.1 show the results that are fairly consistent with the original test.

[Table A.1 about here.]

I then use the expanding training data to run the multinomial regressions and then predict one-year-ahead probability of crashes and jackpots out-of-sample. Starting from 4 years of training sample, the prediction window starts from January 2001 and ends at December 2019. For each month, I form high-minus-low portfolios by sorting stocks based on the predicted crash probabilities into deciles, and then regress either equally weighted or value weighted portfolio returns on CAPM, Fama-French 3-, 4-, 5- and 6-factor models. Table A.2 show the resulting alphas and associated t-statistics estimated using Newey-West standard errors with 12 lags (Newey and West (1986)).

[Table A.2 about here.]

As the table shows, the results from value-weighted portfolios on CAPM, and FF3 and FF4 models are consistent with Jang and Kang (2019). However, they are no longer significant when FF5 and FF6 factors are used, and they are not significant under the equally weighted scheme.

Appendix B. Selected Variable Definitions

ACI = CAPX ratio increase over the previous three periods mean. CAPX ratio is $CAPX/SALE$.

AG = asset growth over the previous year

$Book\_value\_equity$ = $SEQ + TXDITC - Perferred$, preferred is $PSTKRV$, or $PSTKL$, or $PSTK$, whichever is first available.

$Crash\_Risk$ = predicted monthly ex ante probability of stock crash, with log return less than -20%

$FOMC$ = ratio between far out-of-money call option price and the underlying implied forward stock price

$FOMP$ = ratio between far out-of-money put option price and the underlying implied forward stock price

GP = gross profitability, equals $(REVT - COGS)/AT$

$Illiquidity$ = monthly mean of daily absolute return over price times volume of that day, see Amihud (2002).

$Jackpot\_Risk$ = predicted monthly ex ante probability of jackpot, with log return greater than 20%

NOA = $net\_operating\_assets/lag\_AT$

NSI = natural log of changes in adjusted shares

$OSCR$ = $-1.32 - 0.407 \times ASIZE + 6.03 \times TLTA - 1.43 \times WCTA + 0.0757 \times CLCA - 1.72 \times OENEG - 2.37 \times NITA - 1.83 \times FUTL + 0.285 \times INTWO - 0.521 \times CHIN$, O-score, see Ohlson (1980).

ROA = NI/AT

$SMIRK$ = difference between the implied volatility of out-of-money put option and at-the-money call option, see Xing et al. (2010)

$Tang$ = $PPENT/AT$

Table A.1: Replication of Jang and Kang (2019)
The table replicates the multinomial logit regression from Jang and Kang (2019) for sample period 1996 - 2019. Variable definitions follow the paper referenced. Standard errors are clustered at stock and month levels and are included in parentheses.

	Crash		Jackpot	
	Coefficient		Coefficent	
rm12	1.038***	(0.294)	-0.994***	(0.244)
exret12	-0.191***	(0.0509)	-0.211***	(0.0398)
tvol	28.53***	(1.761)	25.20***	(1.178)
tskew	0.0323***	(0.00680)	0.0217***	(0.00766)
size	-0.00731	(0.0121)	-0.154***	(0.0156)
dturn	-0.0238	(0.0494)	-0.315***	(0.0554)
age	-0.0222***	(0.00177)	-0.0149***	(0.00171)
tang	0.121***	(0.0352)	0.119***	(0.0320)
salesg	0.200***	(0.0237)	0.0375	(0.0263)
Constant	-2.968***	(0.273)	-0.511	(0.325)
Observations	965,401		Pseudo R2	0.102
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.2: High-Minus-Low Alphas per Jang and Kang (2019)
The table presents the results from regressing high-minus-low crash risk portfolio returns on various asset pricing factors, following Jang and Kang (2019). Each month, I sort stocks into deciles based on the predicted crash probabilities, and then calculate either equally weighted or value weighted portfolio returns. Then the time series of returns are regressed on time series of asset pricing factors. The sample runs from January 2001 to December 2019. Standard errors are Newey-West standard errors with 12 lags.

	Equal-weighted		Value-weighted	
pricing model	alpha	T-stat	alpha	T-stat
CAPM	-0.348	-0.624	-1.078***	-2.814
FF3	-0.400	-0.920	-1.067***	-3.487
FF4	-0.128	-0.335	-0.866**	-2.635
FF5	0.583	1.156	0.088	0.256
FF6	0.571	1.458	0.081	0.259
Note:			*p<0.1; **p<0.05; ***p<0.01	

References

Abreu, D., Brunnermeier, M. K., 2003. Bubbles and crashes. Econometrica 71, 173–204.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets 5, 31–56.

An, H., Zhang, T., 2013. Stock price synchronicity, crash risk, and institutional investors. Journal of Corporate Finance 21, 1–15.

Andreou, P. C., Antoniou, C., Horton, J., Louca, C., 2016. Corporate governance and firm-specific stock price crashes. European Financial Management 22, 916–956.

Anthony, J. H., 1988. The interrelation of stock and options market trading-volume data. The Journal of Finance 43, 949–964.

Barber, B. M., Huang, X., Odean, T., Schwarz, C., 2020. Attention induced trading and returns: Evidence from robinhood users. Available at SSRN 3715077 .

Barber, B. M., Odean, T., 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. The journal of Finance 55, 773–806.

Barro, R. J., Liao, G. Y., 2020. Rare disaster probability and options pricing. Journal of Financial Economics .

Bates, D. S., 1991. The crash of '87: was it expected? the evidence from options markets. The journal of finance 46, 1009–1044.

Batista, G. E., Prati, R. C., Monard, M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 6, 20–29.

Callen, J. L., Fang, X., 2015. Short interest and stock price crash risk. Journal of Banking & Finance 60, 181–194.

Campbell, J. Y., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. The Journal of Finance 63, 2899–2939.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

Chang, X. S., Chen, Y., Zolotoy, L., 2016. Stock liquidity and stock price crash risk. Journal of Financial and Quantitative Analysis (JFQA), Forthcoming .

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chen, J., Hong, H., Stein, J. C., 2001. Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. Journal of financial Economics 61, 345–381.

Conrad, J., Kapadia, N., Xing, Y., 2014. Death and jackpot: Why do individual investors hold overpriced stocks? Journal of Financial Economics 113, 455–475.

Cooper, M. J., Gulen, H., Schill, M. J., 2008. Asset growth and the cross-section of stock returns. the Journal of Finance 63, 1609–1651.

Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. The Journal of Finance 61, 1605–1643.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990a. Noise trader risk in financial markets. Journal of political Economy 98, 703–738.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990b. Positive feedback investment strategies and destabilizing rational speculation. the Journal of Finance 45, 379–395.

Elliott, G., Timmermann, A., 2016. Forecasting in economics and finance. Annual Review of Economics 8, 81–110.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of .

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., French, K. R., 2020. Comparing cross-section and time-series factor models. The Review of Financial Studies 33, 1891–1926.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of political economy 81, 607–636.

Foucault, T., Sraer, D., Thesmar, D. J., 2011. Individual investors and volatility. The Journal of Finance 66, 1369–1406.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning, vol. 1. Springer series in statistics New York.

Graham, J. R., Kumar, A., 2006. Do dividend clienteles exist? evidence on dividend preferences of retail investors. The Journal of Finance 61, 1305–1336.

Greenwood, R., Nagel, S., 2009. Inexperienced investors and bubbles. Journal of Financial Economics 93, 239–258.

Grinblatt, M., Titman, S., Wermers, R., 1995. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. The American economic review pp. 1088–1105.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220–239.

Han, B., Kumar, A., 2013. Speculative retail trading and asset prices. Journal of Financial and Quantitative Analysis 48, 377–404.

Hirshleifer, D., Hou, K., Teoh, S. H., Zhang, Y., 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hutton, A. P., Marcus, A. J., Tehranian, H., 2009. Opaque financial reports, r2, and crash risk. Journal of financial Economics 94, 67–86.

Jang, J., Kang, J., 2019. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. Journal of Financial Economics 132, 222–247.

Jiang, H., 2010. Institutional investors, intangible information, and the book-to-market effect. Journal of Financial Economics 96, 98–126.

Jin, L., Myers, S. C., 2006. R2 around the world: New theory and new tests. Journal of financial Economics 79, 257–292.

Kelley, E. K., Tetlock, P. C., 2017. Retail short selling and stock prices. The Review of Financial Studies 30, 801–834.

Kelly, B., Jiang, H., 2014. Tail risk and asset prices. The Review of Financial Studies 27, 2841–2871.

Kim, J.-B., Li, Y., Zhang, L., 2011. Corporate tax avoidance and stock price crash risk: Firm-level analysis. Journal of Financial Economics 100, 639–662.

Kim, Y., Li, H., Li, S., 2014. Corporate social responsibility and stock price crash risk. Journal of Banking & Finance 43, 1–13.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political analysis 9, 137–163.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics 45, 221–247.

Merton, R. C., 1973. An intertemporal capital asset pricing model. Econometrica: Journal of the Econometric Society pp. 867–887.

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics 108, 1–28.

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109–131.

Pan, J., 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. Journal of financial economics 63, 3–50.

Petersen, M. A., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. The Review of Financial Studies 22, 435–480.

Ripley, B. D., 1996. Pattern recognition and neural networks. Cambridge university press.

Ritter, J. R., 1991. The long-run performance of initial public offerings. The journal of finance 46, 3–27.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Seliya, N., Khoshgoftaar, T. M., Van Hulse, J., 2009. A study on the relationships of classifier performance metrics. In: *2009 21st IEEE international conference on tools with artificial intelligence*, IEEE, pp. 59–66.

Titman, S., Wei, K. C. J., Xie, F., 2004. Capital Investments and Stock Returns. Journal of Financial and Quantitative Analysis 39, 677–700.

Welch, I., 2020. Retail raw: Wisdom of the robinhood crowd and the covid crisis. Tech. rep., National Bureau of Economic Research.

Wermers, R., 1999. Mutual fund herding and the impact on stock prices. the Journal of Finance 54, 581–622.

Xing, Y., Zhang, X., Zhao, R., 2010. What does the individual option volatility smirk tell us about future equity returns? Journal of Financial and Quantitative Analysis pp. 641–662.

Yan, S., 2011. Jump risk, stock returns, and slope of implied volatility smile. Journal of Financial Economics 99, 216–233.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67, 301–320.