Predicting IPO Performance from Prospectus Sentiment

Senhao Ni

College of Art and Science, Boston University, Boston, USA

Corresponding author: eleven@bu.edu

Abstract. With the development of society, how to apply artificial intelligence in finance becomes a hot topic. IPO performance is the most significant part of these areas. Therefore, this paper aims to use sentiment analysis to predict the IPO performances in the Hong Kong stock market. In specific, 6 machine learning models are trained, namely Random Forests, Decision Tree, Naïve Bayes, Logistic Regression, LightGBM and a stack model to predict the direction of Hong Kong stocks IPO performances (positive, negative, or little change) in 3, 5,10, 20 and 30 days after their IPOs. This paper will compare all the models with a baseline model which generates random guesses for the direction of IPO performances to conclude about which model works better and which one is more predictable. This paper will show that instead of logistic regression, random forest performs relatively the best across all Ys, and this may be due to the different sample size.

Keywords: multifactorial prediction; linear regression; machine learning model.

1. Introduction

In recent years, there have been increasing issues related to IPOs in Hong Kong. According to Sina Finance, the rate of fall on the first trading day of Hong Kong stocks has reached 74% in 2021, which together with the spread of omnicorn in Hong Kong and the bigger fall in its local stock market indices, has resulted in the slump of IPO deals in Hong Kong in 2022 according to CNBC news. Moreover, as is known to all, the Russian-Ukrainian conflicts and the rate hikes from the Fed are all shaking the financial markets even more at this point in time. Living and investing during such a fast-changing time, this paper may dig as many "certainties" as possible to stay profitable in investing in the IPO market, and stock prospectus as a "certain" document that will always be published before every IPO, is now worthy of more attention.

According research, sentiment analysis can be used to predict the IPO performances in the Hong Kong stock market. In specific, there will be training 6 machine learning models, namely Random Forests, Decision Tree, Naïve Bayes, Logistic Regression, LightGBM and a stack model to predict the direction of Hong Kong stocks IPO performances (positive, negative, or little change) in 3, 5,10, 20 and 30 days after their IPOs. Then, all the models will be compared with a baseline model which generates random guesses for the direction of IPO performances to conclude which model works better and which one is more predictable.

2. Organization of the Text

2.1 Data Preprocessing

2.1.1 Data Collection

In this project, this paper uses data obtained from the website of Hong Kong Stock Exchange. This paper collected prospectus documents under search options listed below:

Search Type: Current Securities.

Headline Category and Document Type: Headline Category; Listing Documents→Offer for Subscription.

From time and To time: intervals not exceeding 12 months.

News Title: global offering.

Subsequently this paper used web-crawling techniques to mass-download documents, collected stock codes associated with these documents, and excluded repeated files, which in detail are: submissions with same stock code but respectively single and multiple files, and submissions of the

same stock at different times. For the latter case, this paper chose the document submitted closer to the listing date.

The information of each stock on listing date, IPO price, percentage change of price in different days after listing (1, 3, 5, 10, 20 and 30), and IPO Lead Underwriter are downloaded from website Wind. With these data, this paper excluded prospectus documents of which the stock code cannot be traced. Search period is from 2011 to 2022, and for 2022 this paper included the first 9 prospectuses. Total data size is 815 rows.

2.1.2Text Extraction

-Tokenization

This paper used the python package pdfplumber to extract text from pdf files in page sequence. In the extracted texts this paper only retains letters and set aside figures to better evaluate the verbal skills displayed in these documents. Hence in the string of text of an extracted page, if a character is a letter, it is added to the end of the last token listed. Otherwise it suggests a break and when the next immediate letter is observed, a new token is created. By this means this paper could collect all the letters split by non-alphabetical characters, and by a rough inspection it seems that all such tokens collected are individual words.

-No Stemming or Lemmatization

Collected tokens did not undergo stemming or lemmatization process as the dictionary this paper used contains derivatives.

2.1.3 Text Scoring by Frequency

To replicate experiment steps of one previous investigation, this paper utilized Loughran-McDonald dictionaryfor word scoring. this paper counted word length and character length, and proportions of all words under categories “Negative”, “Positive”, “Strong Modal Words”, “Weak Modal Words”, “Uncertainty Words” and “Litigious Words”. The original report also contains categories “Moderate Modal Words” and “Complex Words”; these two categories were treated specially:

“Moderate Modal Words”: This category no longer exists in the up-to-date version of dictionary, and hence is not searched for. For compensation, an additional feature “Constraining” is extracted from the dictionary and added to our data.

“Complex Words”: According to experiment steps of one previous investigation, words that are made up of two or more components are defined, where “components” is an ambiguous term. Hence, according to Loughran-McDonald dictionary, this paper executes the standard of the R package on sentiment analysis that use the same dictionary, and define a complex word as having more than 3 vowels (“a”, “e”, “i”, “o” and “u”).

The difference between the original experiment and our research is that this paper uses an indicator similar to term frequency (total word count of all words of a type over word length), instead of simple word count, out of concern that writers might produce lengthy prospectus, boost words with positive connotations and tort our judgment on the qualities of their writings.

2.1.4 Stock Data for Prediction

This paper collated percentage change in stock price 1, 3, 5, 10, 20, 30 days after listing and predicted the trends of these six changes. Stocks with prices that fluctuate less than 0.5% are labeled 0; those with prices increased at least 0.5% are labeled 1 and the rest -1.

Table1. Part of Training Data

Stock Code	Word_ Length	Char_ Length	Negative	Positive	Strong_Modal	Weak_Modal	Uncertainty	Litigious	Constraining	Complex	Day 1	Day 3	Day 5
1549. HK	-0.74	-1.23	-0.75	-0.34	0.40	0.08	0.74	-0.40	-0.30	-1.40	1	-1	-1

1431. HK	0.21	-0.27	-0.26	0.84	0.20	-0.16	-0.61	-1.04	-0.36	-0.87	-1	0	-1
1257. HK	-0.30	1.13	-0.02	-1.45	-1.12	-0.79	-1.35	-1.94	-1.55	2.36	-1	0	1
1231. HK	-0.11	0.33	-1.30	-0.02	1.05	-0.41	-0.01	-1.13	-0.85	0.59	-1	-1	-1
1815. HK	0.11	-0.32	-0.90	0.79	1.16	-0.46	-0.97	-1.29	-1.46	-1.15	0	-1	-1
...	…	…	…	…	…	…	…	…	…	…	…	…	…
2250. HK	-0.32	0.15	-0.14	-0.14	-0.52	-0.83	1.91	-0.11	-1.49	0.58	1	-1	-1
6068. HK	0.28	0.19	-0.49	0.07	0.22	0.65	-0.35	2.01	1.17	-0.43	-1	1	1
2033, HK	0.11	0.20	-1.20	-0.34	0.10	-0.48	-0.45	-1.22	-0.199	-1.05	1	0	1
9996. HK	-0.16	0.48	1.33	0.58	-0.72	1.64	0.63	-0.47	-0.45	0.95	1	1	1
1300. HK	-0.90	-1.06	-0.45	-0.49	0.54	-0.58	0.83	-0.21	-0.57	-0.60	1	1	1

Since Day 1 price is unstable, this paper only studied the rest of the predictions, i.e. meaningful comparisons were made only on stock prices of days 3, 5, 10, 20 and 30.

2.1.5 Train-test Split

This paper arbitrarily chose 100 rows as the test set; the remaining 715 rows are the training set.

2.2 Features Engineering

To select/create the most effective features, this paper first looks at the correlation matrix, trying to identify high importance features for any Y in different time period.

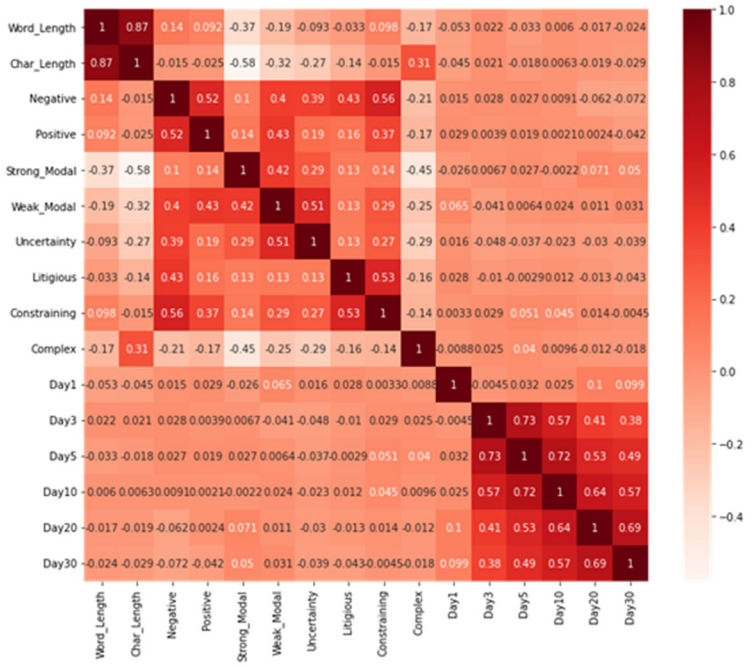

Fig 1. Feature Correlations

However, no features show high importance for Y. Thinking intuitively, this paper supposes that some mixtures of sentiments may be more informative, for example, the stock prices may be positive if its prospectus conveys not only strong certainty (Strong_Modal) but also positive view on their

future development(Positive), and vice versa. Therefore, this paper is going to find proper interaction features that may improve our model performance.

Firstly, this paper created 56 features, including the original 10 features, 45 one-by-one interaction terms, and 1 interception term.

Secondly, according to the paper, Logistic Regression performs relatively the best, so this paper is going to tune a Logistic Regression as a means for further forward feature selection tuning. Also, Logistics Regression requires little/no correlations among all the variables, so it is considered that removing one of Word_length and Char_length as they are highly correlated. However, it is uncertain about which one to drop, so this paper is going to keep it temporarily. After tuning the Logistic Regression classifier based on the original features, this paper will show the following best parameters and its prediction accuracy:

Table 2. Parameters and prediction accuracy

Days	Best parameter	Accuracy
Day3	best_params_ (‘C’ : 0.01,‘solver’: ‘newton-cg’)	0. 49
Day5	best params_ (‘C’: 0.01,‘solver’: ‘newton-cg’) acc 0. 48	0. 48
Day10	best params_ (‘C’: 0.01,‘solver’: ‘newton-cg’) acc 0. 48	0.54
Day20	best params (‘C’: 0. 05 ‘solver’: ‘newton-cg’)	0.53
Day30	*best params_ (‘c’: 0.01,* ‘solve’ :‘newton-cg’)	0.53

Noted that in most cases, C=0.01 with ‘newton-cg’ as solver was chosen by the GridSearchCV, therefore, this paper is going to use the estimator with these parameters (others use default), together with *SequentialFeatureSelector* to do *forward feature selection*. Meanwhile, this paper is using the return of Day30 as the dependent variable because when searching for return data, Day30 is missing the least amount of data.

Thirdly, to tune the forward-feature-selector, this paper used the number of features ranging from 2 to 54 (total = 56). The overall distribution looks like this:

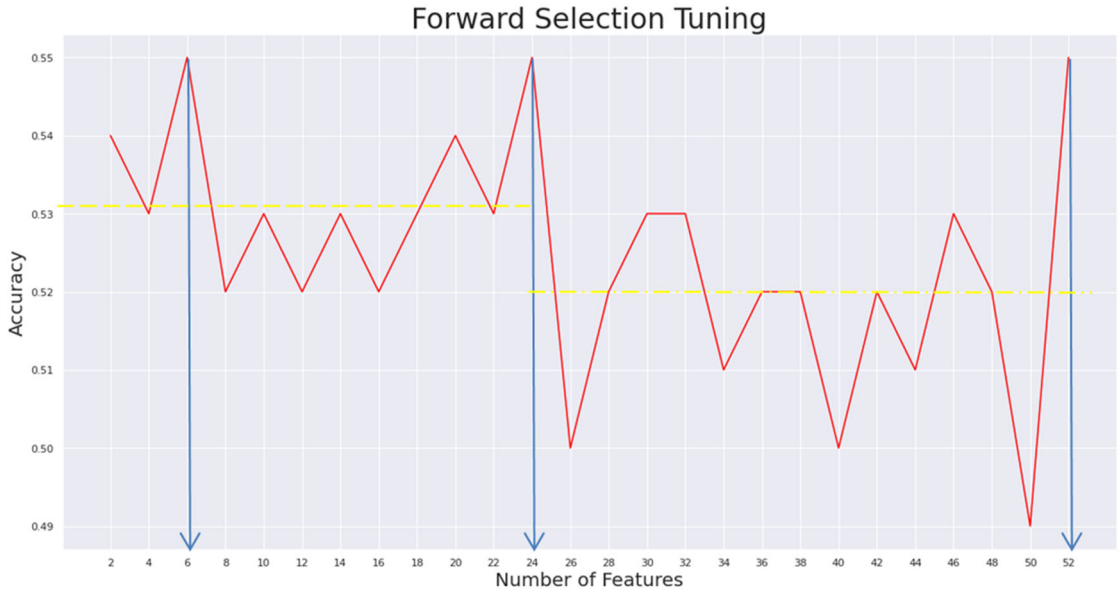

Fig 2. Accuracies against Feature Sizes

From the plot, this paper finds that number of features = 54 gives the highest accuracy score = 0.56 (not in the graph), however, there is no obvious trends for accuracy increases with the number of

features, instead, it seems like smaller number of features has a slightly higher average level of accuracy (in the yellow dotted line), therefore, this paper tends to try using the least amount of features that can give us a relatively high accuracy >= 0.55, which is the accuracy that the model with the number of features=6, 24, 52 gave us (pointed by the blue arrow). To achieve this, this paper is going to drop ALL features that have been dropped by model(n_features_to_select=6) and model(n_features_to_select=24), and add only the best combination of the features that has been added by ALL the 3 best models.

After summarizing, dropping and adding features, this paper finds 5 models with only 4 features (2 from original, 2 from interaction) that all give prediction accuracy = 0.55.

GridSearch again this paper gets the comparison table of accuracy:

Litigious + Complex + Word Length * Strong Modal + Weak Modal * Constraining new feature acc: 0.55

Litigious + Complex + Word Length * Strong Modal + Litigious * Constraining new feature acc: 0.55

Litigious + Complex + Char Length * Negative + Weak Modal * Constraining new feature acc: 0.55

Litigious + Complex + Negative * Complex + Weak Modal * Constraining new feature acc: 0.55

Litigious + Complex + Weak Modal * Constraining + Litigious * Constraining new feature acc: 0.55

This paper is going to go with m3 because it improves all the return predictions the most. Our selected features are Litigious, Complex, Char_Length * Negative, and Weak_Modal * Constraining. However, this paper will still be trying different features for different methods in the following sections from our bag of features (totalling 56) when m1-m5 is not improving the model prediction performance.

Table 3. Feature Engineering (Continued)

	before feature selection	M1	M2	M3	M4	M5
Day3	0.49	0.49	0.49	0.50	0.50	0.49
Day5	0.48	0.49	0.50	0.50	0.50	0.50
Day10	0.54	0.57	0.57	0.57	0.56	0.57
Day20	0.53	0.53	0.54	0.53	0.53	0.53
Day30	0.53	0.55	0.54	0.55	0.55	0.56

2.3 Methods

2.3.1 Baseline Model: Random Assignment

A natural method of guessing stock trends is by proportion. This paper devised two assignment schemes: weighted random assignment and non-weighted random assignment. For the first approach, for each indicator to predict, this paper measured fractions of occurrences of -1, 0 and 1 in the training set, and for the 100-row test set, this paper generated 100 randomly sampled numbers within -1, 0 and 1 by the weights measured. For the non-weighted approach, this paper simply randomly generates labels with equal weight. After comparison this paper chose weighted assignment as the baseline model.

2.3.2 Random Forests

The previous research shows that random forests outperform the alternative classic machine learning algorithms in terms of mean and median predictive accuracy when predicting. This paper did a grid search to tune the following parameters of Random Forest

Classifier: maximum depth, maximal number of features, minimal number of samples leaf, minimal samples split and number of estimators. The predicted target is a categorical variable representing three trends which are up (1), down (-1) and unchanged (0) respectively, compared to

the stock issue date price. After tuning, this paper obtains a set of parameters for the best models corresponding to the time points this paper needs to predict.

2.3.3 Decision Tree

The most important parameter in the decision tree model is the maximum depth of the tree. In general, the deeper the tree grows, the more complex the model will become because this paper will have more splits and it captures more information about the data. In order to get a proper range of maximum depths for subsequent grid search, this paper records the training accuracy and test accuracy inside a relatively large loop. This paper finds when depth is bigger than 15, there will be overfitting. So this paper ends up with max_depth in range (3, 15), leave the other parameters as default, and then use grid search to find the optimal tree depth.

The performance of a tree can be further increased by pruning. It involves removing the branches that make use of features having low importance. After checking the feature importance of our model, this paper finds that there are three features with zero importance: Strong_Modal,Weak_Modal,Complex. This paper drop these features and get better accuracy results after tuning.

2.3.4 Naive Bayes

The BernoulliNB model of Naive Bayes algorithm is tuned in this stage, and there is only one parameter alpha, which is used to control the complexity of the model. The working principle of alpha is that the algorithm adds so many virtual data points to the data. These points take positive values for all features, which can smooth the statistical data. The larger the alpha, the stronger the smoothing, and the lower the complexity of the model. Here, this paper tuned the alpha through the grid search method and find the optimal parameters, the best parameter for Day30 is 700, and based on the feature's selection above, this paper retrained the model by all five feature collections, but it shows there is no difference on accuracy to do feature selection. However, after retraining the model of newly-selected features, the best alpha is 100.

2.3.5 Logistic Regression

According to Kambria[10],Logistic regression is a powerful machine learning algorithm that utilizes a sigmoid function and works best on binary classification problems. Unlike other machine learning algorithms that this paper has tried, logistics regression is the only parametric model that has a finite number of coefficients (weights) for each variable, and it has stricter assumption requirements. Now, this paper will check the assumptions claimed by Complete Dissertation [11] one by one for the data (after feature selection in section4):

Logistic regression requires the observations to be independent of each other.

Logistic regression requires there to be little or no multicollinearity among the independent variables.

Logistic regression assumes linearity of independent variables and log odds.

Logistic regression typically requires a large sample size.

Table 4. Correlation Tables for Engineered Features

	Litigious	Complex	Char_Length * Negative	Weak_Modal * Constraining
Litigious	1.00	-0.16	0.19	0.09
Complex	-0.16	1.000	-0.04	-0.39
Char_Length • Negative	0.19	-0.04	1.00	0.01
Weak_Modal• Constraining	0.09	-0.39	0.01	1.00

Because prospectuses are from different industries across different time ranges, this paper assumes criterion 1 is satisfied. According to figure 8, engineered features do not display strong or moderate correlations, and hence criterion 2 is satisfied. Furthermore, figure 3(plotted in the following page) indicates that criterion 3 is satisfied. To save space, only Day30 log odds linearity plots are displayed

below. Plots for Other Ys can be seen in codes. In figure 3, this paper looks for S-shaped curves with flat top and bottom with an increasing or decreasing middle, and it seems like our selected features meet the conditions. Same for another Ys.

Unfortunately for the last criterion, due to data limitations, this paper does not have a sample as large as the paper, and this may limit the performance of our logistics regression model.

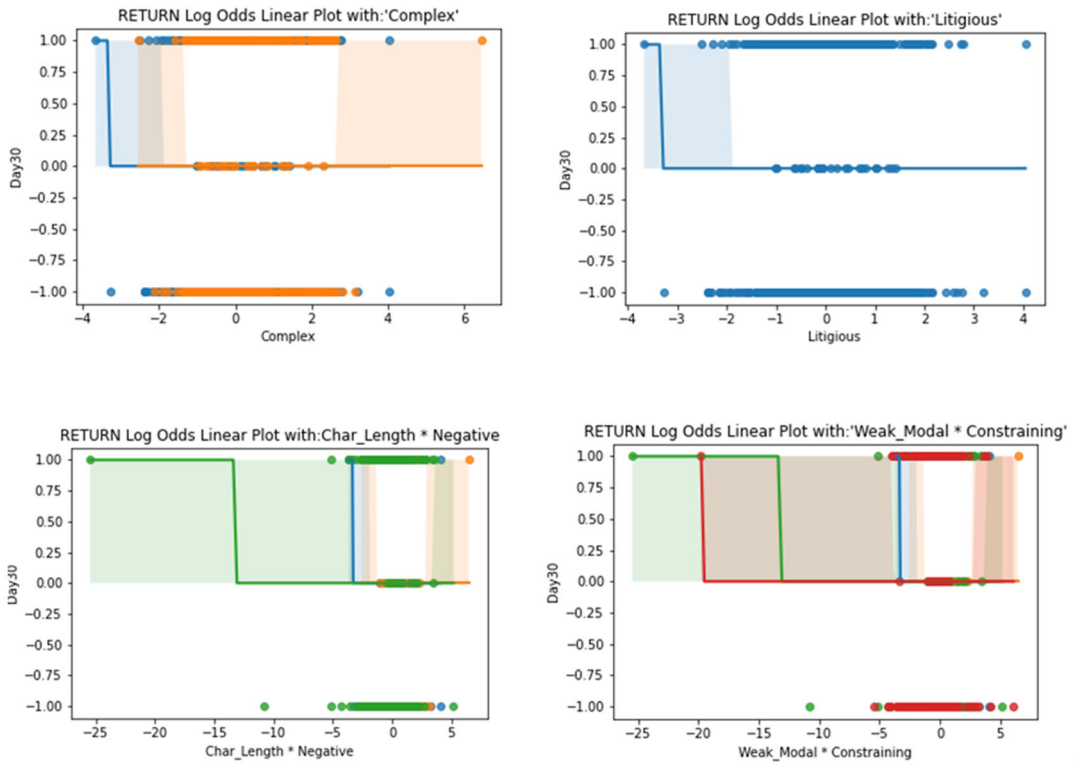

Fig 3. Feature Log Odds Linearity Plots of Day30

After checking the assumptions, sklearn.linear_model.LogisticRegression is used for model building. The only hyper-parameter in it is "C", which needs to be tuned and this parameter represents the inverse of regularization strength, the lower the stronger the regularization. The "solver" will also be tuned, as their algorithms are not clear to us. "penalty" is not being tuned as "l2" is the only available regularization for most solvers that are able to work for multi-class classification

2.3.6 Light GBM

A five-fold grid search with cross-validation is tried with maximum depth is tuned from 1 to 10, and number of estimators from 5 to 200 with interval 5. Search grid hence composes of 400 combinations. For input features, both original features (10 in total) and those after feature selection (54 in total) are tried.

2.3.7 Stacking

Stacking is an ensemble learning technique to combine multiple classification models via a meta-classifier. In our project, this paper adds all the above basic models with their respective best parameters on level 0 of stacking. And this paper chooses logistic regression model as a meta-classifier on Level 1. The meta-classifier is fitted based on the outputs -- meta-features -- of the individual classification models in the ensemble.

2.4 Results

2.4.1 Baseline Model: Random Assignment

With weighted random assignment, prediction accuracy typically varies between 30% and 50%, and accuracies for Day 20 and Day 30 often exceed 40%. With non-weighted random assignment, prediction accuracy typically varies between 25% and 50%.

The relatively better performance of the baseline model in the prediction of Days 20 and 30 might be caused by convergence of stock trends. As longer days pass since a stock is listed, random price fluctuations are more strongly rectified by determinant factors of stock price as these factors gradually take effect. Hence in latter days test results are more predictable from training results.

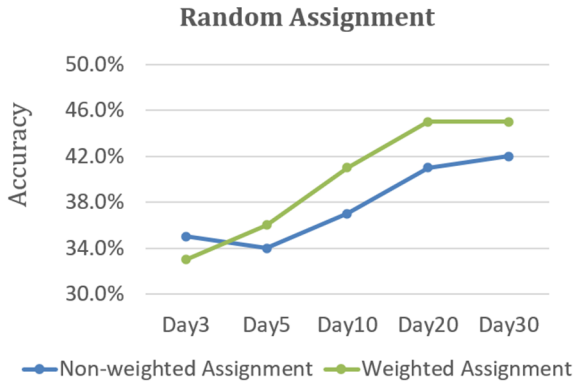

Figure 4. Sample Output by Random Assignment.

Numbers in Blue Line: 35%, 34%, 37%, 41%, 42%.

Numbers in Green Line: 33%, 36%, 41%, 45%, 45%. (Final Baseline Model)

2.4.2 Random Forests

5-Fold cross validation was applied on the training data to fully use the dataset. Focused on the accuracy_score of the predicted results, the prediction performance varies from 38% to 57%, and the overall the accuracy improved over time. Though the model outperformed the baseline, the prediction accuracy in D3 and D5 still remain below 50%, which is not very optimal.

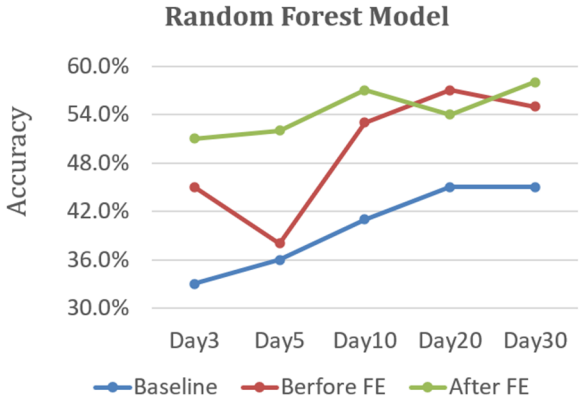

Fig 5. Test Accuracy of Random Forest Model

After introducing the interaction terms as new features, the performance did improve across the time, the accuracy with feature engineering varies from 51% to 58%. The new model with feature engineering consistently outperformed the baseline model and it improves more for shorter-term prediction scenarios (e.g., day 3 & day 5), which is exactly what this paper expects.

2.4.3 Decision Tree

This paper applies 5-fold cross validation on the training dataset to get the best parameter for each case and use accuracy score to evaluate. We could see that the decision tree model has evident better performance than baseline for all cases. And our pruning process also improves the model performance on day 5, day 20 and day 30. Compared with baseline, the final model's best improvement is 16% over baseline on day 3, and its worst is 10% over baseline on day 20 and 30.

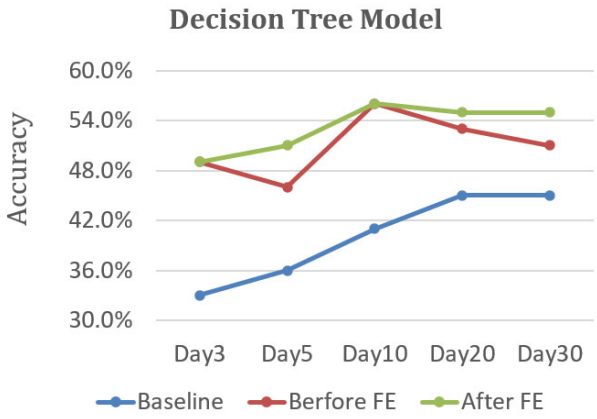

Fig 6. Decision Tree Test Accuracies

2.4.4 Naive Bayes

The forward feature selection collection makes no difference in the accuracy of the Naive Bayes model, after training the model, the accuracy for Day 1, Day3, Day5, Day10, Day20, Day30 is 0.53, 0.5, 0.49, 0.57, 0.53, 0.54 respectively, from which can find that the Naive Bayes model exceeds the Baseline model to some extent.

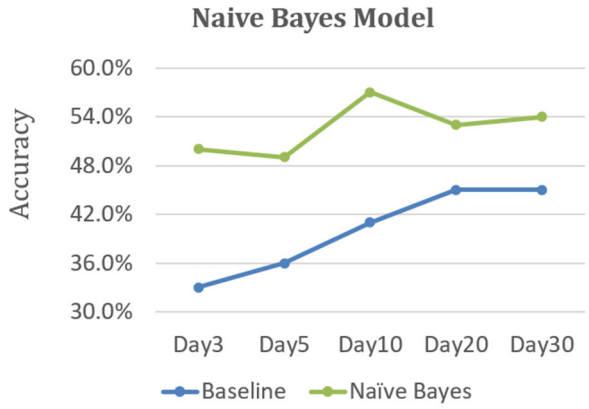

Fig 7. Naive Bayes Test Accuracies

2.4.5 Logistic Regression

This paper applied 5-fold cross validation on both the original training dataset and the after-selection training dataset. This paper also plot the prediction accuracy for the test sets, and the (baseline accuracy, before FE accuracy, after FE accuracy) for Day 3 is (0.33,0.49,0.5), for Day5 is (0.36,0.48,0.5), for Day10 is (0.41, 0.54, 0.57), for Day20 is (0.45, 0.53, 0.53), for Day30 is (0.45, 0.53, 0.55), on average the logistics regression model (after FE) has improved the prediction accuracy by 0.13.

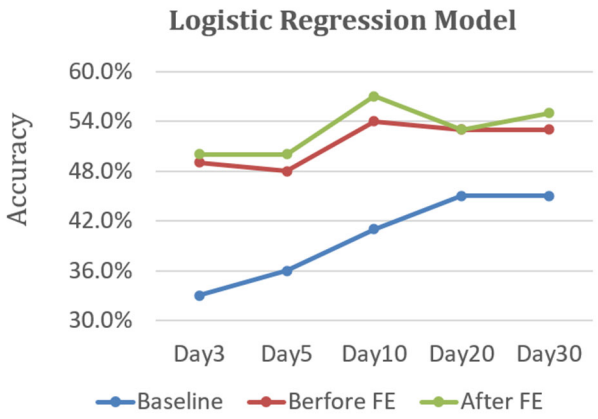

Fig 8. Logistic Regression Test Accuracies

2.4.6 Light GBM

Simple models are often preferred, with maximum depth 1 and number of estimators 5 occurring in the optimal models for predicting stock price change on most days, with or without feature selection. The optimal model in these two cases achieves identical test accuracy 57%.

In retrospect, a possible explanation for such an outcome might be that: using accuracy as a scoring criterion might cause labels that occur less frequently to be under-represented. In both train and test data, largest number of occurrences of 0 appear in Day 1, which are all less than ⅙ of the total number of rows and monotonically decrease as days pass. Thus, Light GBM method might prefer to treat this classification problem as a binary one, making at most one cut to the input space.

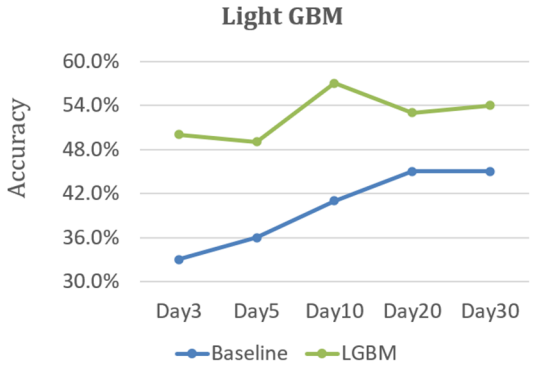

Fig 9. Light GBM Test Accuracies

Green Line Signifies Performance without Feature Engineering

2.4.7 Stacking

After stacked all the above models with their respective best parameters, this paper shows that the accuracy performance is obviously better than the baseline model, varying from 0.47 to 0.58. However, it only outperforms the single optimal model on day 10. So it is clearly show that stacking is not very helpful on our dataset, because our's is relatively small and doesn't have many features.

2.5 Discussions

Benchmark to Define Predictive Capacity

The study uses ROC-AUC to score models, and hence its standards are not directly comparable to ours. Our method of weighted random assignment can be considered reasonable in generating baseline prediction outcomes. As illustrated above, all our models outperform baseline outcomes, which demonstrates that modeling is effective.

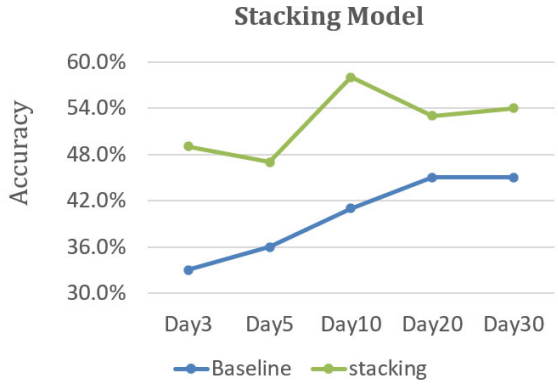

Fig 10. Test Accuracy after Stacking

Comparing Models in Referred Report

According to report , logistic regression produces superior outcomes for all the predicted stock price indicators. In our experiments, random forest and decision tree can be considered good methods, as their prediction accuracies for Day10, Day20 and Day30 do not fall below 54%. This is probably due to their simple mechanisms, which might work well when less data is provided. In comparison, logistic regression requires extensive assumption and relatively larger dataset, of which ideal values are difficult to obtain. Additionally, the effect of stacking seems limited. A graph with all prediction accuracies combined is provided below for readers' reference.

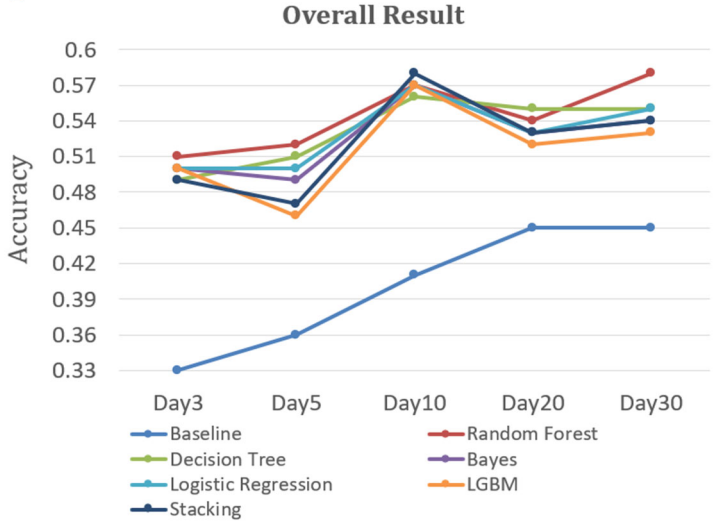

Fig 11. Overall Results

Simple models being more appropriate for small databases might also explain the low effect of feature engineering on Light GBM and Naive Bayes models: they have simple processes; complicating input features might be counterproductive.

Miscellaneous: ROC-AUC as Scoring Criterion

Although as mentioned in the introduction of LightGBM scores, our classification problem can be viewed as a binary one, this paper subjectively disallowed such interpretation as for most of the indicators that have been predicted, namely days 1, 3, 5, 10, occurrences of 0 in both train and test set, despite being low, do not fall below 5%, with 0 on day 20 taking 33 out of 715 rows in training set and 9 out of 100 rows in test data. As such did not reuse the scoring criteria ROC-AUC from the original report . We did try to construct a multinomial roc-auc scoring method by first aggregating all false positive rates, then interpolating all ROC curves between these, and finally averaging them and compute AUC. However this was not effective in distinguishing models with strong predictive capacity from the rest since all models' ROC-AUC score passed 0.7.As our experiments run on relatively few documents, they can be repeated with more data inputs to verify our discoveries.

3. Literature References

3.1 Related Work Overview

The impacts of public textual sentiment on stock returns have been studied in some detail, among which the sentiment analysis towards the prospectuses is the extensive one. The prospectus is a legal document that explains the investment/security and the offering in detail. It includes information on the firm (or fund), its finances, management, and other factors that may assist investors in making an informed decision.

Ferris et al. acquire a number of key conclusions based on a textual examination of initial public offering prospectuses, including the relationship between prospectus conservatism, IPO price, and subsequent operational and stock return performance. Prospectus conservatism is positively connected to underpricing according to their findings. In addition, according to previous research, prospectus conservatism may predict the firm's post-IPO operating success for non-technology IPOs as well.

Another According to research, this research made attempts to look at the amount of ambiguous language in the prospectus as well as its closeness to previously filed registration statements to see whether textual factors may help solve the IPO underpricing mystery. According to Paulus et al., the result indicates that the prospectus' ambiguous language does not reflect the issuer's expectations about the company's future prospects, but textual analysis can actually help to explain underpricing of US REIT IPOs.

Forward-looking statements (FLSs) provide useful information in applications such as stock price forecasting. According to previous research FLSs are found in the Management Discussion & Analysis portion of prospectuses for initial public offerings, and they provide prospective information regarding the company's future development and performance. Topics, feelings, readability, semantic similarity, and general text characteristics are among the linguistic features which Tao, Deokar, and Deshmukh engineered from FLSs. As suggested by their research conclusion, FLS traits are more predictive of pre-IPO valuation prediction than post-IPO valuation prediction.

3.2 Same Prediction Problem

Ly and Nguyen's paper - Do Words Matter: Predicting IPO Performance from

Prospectus Sentimenthave also examined the sentiment on prospectuses in order to estimate the price movement of an initial public offering pricing. They proposed a new methodology that employs sentiment analysis to anticipate an IPO's price movement throughout the first three, five, ten, twenty, and thirty days. The features they looked at were file size, character count, complex words count, negative words count, positive words count, strong modality words count, moderate modality words count, weak modality words count, uncertainty words count, and litigious words count to see how big an effect words in a prospectus can predict the fate of an IPO in the short term. Through establishing a baseline model and the following models, this paper can show that only the logistic regression model optimally outperformed the baseline model throughout the time, while other models consistently performed worse than the base model or outperformed it only at day 30.

On Day 30, the logistics regression model had a peak accuracy of +9.6% from baseline, and the performance increased as the date got further from the offer date. The author suspects that this is due to investors being more sensible and having more time to evaluate financial information such as the prospectus.

4. Summary

Through examination on input data and our prediction outcomes, this paper has discovered that stock price trends are more predictable typically after ten days of listing, where some models perform worse in predicting stock trends after ten days than in predicting those before. Through careful

research, this paper has discovered that instead of logistic regression, random forest performs relatively the best across all Ys, and this may be due to the different sample size.

References

[1] https://finance.sina.com.cn/stock/hkstock/ggipo/2021-12-31/doc-ikyamrmz2237292.shtml.

[2] https://www.cnbc.com/2022/04/05/ipos-in-china-hong-kong-down-amid-omicron-surge-stock-volatility.html.

[3] S.P. Ferris, G. Hao and S.M. Liao (2013), The Effect of Issuer Conservatism on IPO Pricing and Performance, Review of Finance, Volume 17, Issue 3, Pages 933-1027.

[4] N.M. Paulus, M. Koelbl and W. Schaefers (2021), Can textual analysis solve the underpricing puzzle? A US REIT study, JOURNAL OF PROPERTY INVESTMENT & FINANCE, DOI: 10.1108/JPIF-06-2021-0052.

[5] J. Tao, A. V. Deokar and A. Deshmukh (2018) Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach, Journal of Business Analytics, 1:1, 54-70, DOI: 10.1080/2573234X.2018.1507604.

[6] T. H. Ly and K. Nguyen (2020), Do Words Matter: Predicting IPO Performance from Prospectus Sentiment, 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pages. 307-310, DOI: 10.1109/ICSC.2020.00061.

[7] Information from: https://www1.hkexnews.hk/search/titlesearch.xhtml.

[8] Information from: https://sraf.nd.edu/loughranmcdonald-master-dictionary/.

[9] Information from: https://sraf.nd.edu/loughranmcdonald-master-dictionary/.

[10] Information from: https://kambria.io/blog/logistic-regression-for-machine-learning/.

[11] Information from: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/.