A New Machine Learning Framework for the Prediction of Companies' Business Success

Yiheng An (805640602), Xinyi Zhang (805641673), and Beier Cao (605638822)

***Abstract*— As the many shocking value-destroying events happened over the outbreak of COVID-19, it is more important for investors to make good decisions under greater uncertainty. In this report, we introduced the exchange rate data and Google Trends web search data to an innovative machine learning framework for the prediction of companies' business success. We detected and addressed the problem of imbalanced labels and found that KNN, RF and LR in our system showed robust performance over both large and small datasets. In this report, we will also demonstrate how exchange rate related features and web search related features add value to the traditional purely company-based framework, as well as discuss potential user scenarios of powerful classifiers in our framework.**

***Index Terms*—Machine learning, predictive modelling, business success.**

I. Introduction

Due to the impact of COVID-19, the future of companies has become more elusive for many people. Even for sophisticated investors, they are surprised to see a series of famous multinational corporations and firms released their involuntary bankruptcy petitions in a row, including NPC International, Virgin Atlantic and so on. Given these shocking value-destroying events nowadays, it has become increasingly crucial for investors to make wiser decisions. Fortunately, the innovation of algorithm and the release of computing power has provided new possibilities. Classifiers, a set of machine learning methods that tell us one specific status from finite possibilities, has become more and more developed and easy to apply. By labelling companies' status, many powerful supervised models can be involved in building the connection between the predictive information we have (features) and the companies' future (targets).

In this project, we try to design a robust predictive framework of companies' business success. For each company, we define the business success as an event of M&A (Merger and Acquisition) or IPO (Initial Public Offering). Many advanced techniques in feature extraction, feature engineering, predictive modelling are involved. Based on the framework we propose, we will implement different models and evaluate their performance on both validation dataset and testing dataset. Also, we will show whether different types of features add value to the predictive performance. Finally, we will train our machine learning system over smaller sets of data to explore the robustness of it.

II. Statement of Previous work

This section covers previous research articles related to our topic. Our initial, fundamental thoughts and datasets are based on the report "Machine Learning Prediction of Companies' Business Success", written by Chenchen Pan, Yuan Gao and Yuzi Luo [1]. In it, they tried to apply different machine learning models, such as K-Nearest Neighbors (KNN), Logistic Regression (LR) and Random Forests (RF) model, to build a machine learning system that can help to predict companies' success in failure. By using F1 score, they concluded that KNN model has a better performance which achieves an F1-score of 44.45% and accuracy of 73.70%. We believe there is space to further explore and refine this original system from the perspective of both design and performance.

In Bento's paper[2], the author built a predictive system applying Random Forests to classify success or failure start-ups based on M&A's financial reports. In the binary classifier, he created the Confusion Matrix exams companies' future and indicates that the True Positive Rate (TPR) is 94.1% and the False Positive Rate is 7.8%.

Afolabi, Ifunaya and Moses, from Department of Computer and Information Sciences in Covenant University, did a similar research on the elements that may affect a company's success in their journal[3]. They designed a diagnosis system for business prediction and recommendation. Their methodology concerns the correlation analysis for data processing combines with Naïve Bayes method and J48 classification algorithm. The model they created has a prediction accuracy up to 70% in predict how long a business will last. This journal offers us an idea of the research procedure and inspires us to consider features related to government policies and macroeconomics.

In the article, Machine Learning based Business Forecasting, Singh and his team members focused on the order to build forecasting models based on sales data from internal airline companies [4]. They applied Gaussian process, linear regression, multilayer perceptron and so on and conducted evaluation through metrics such as mean absolute error (MAE) and root mean squared error (RMSE). This work inspires us to think about efficient and effective ways to evaluate the performance of our system and its robustness.

TABLE I
FEATURE EXTRACTION AND LABELING

Name	**Description**
Company-Related Features/Label	
Country_code	The country code given to each company
Funding_total-usd	The total amount of funding in all rounds of investments
Funding_rounds	The total number of funding rounds
Invest_duration	The number of weeks between the first time and the last time raised money
Label	The operation status of the company (0 = closed or operating, 1 = IPO or acquired)
Exchange Rate Features	
Sd_overyear	The standard deviation of exchange rate (Gold/Target currency) over the whole-time window
Ex_sd_5years	The standard deviation of exchange rate (Gold/Target currency) over 5 years since the company founded
Ex_mean_5years	The mean of exchange rate (Gold/Target currency) over 5 years since the company founded
Web Search Features	
IPO_mean_5years	The mean of monthly web search scores of IPO over 5 years since the company founded
IPO_rate_5years	The growth rate of monthly web search scores of IPO over 5 years since the company founded
M&A_mean_5years	The mean of monthly web search scores of M&A over 5 years since the company founded
M&A_rate_5years	The growth of monthly web search scores of M&A over 5 years since the company founded

III. TASK DESCRIPTION

In this project, we choose to apply various machine learning methods to deal with a classification problem. The goal is essentially to classify the status of companies based on historic data we obtained. So, the first task is to set up a reasonable standard which defines the label of business success or not for each company.

Besides, we need to build competing supervised classifiers and evaluate their performance. Our evaluation would focus on the robustness of models in generalization and the value of the exchange rate features and web search features in this application.

Moreover, based on the performance of each algorithm, we hope to summarize possible user scenarios for powerful models and make our project more practical and useful.

IV. MAJOR CHALLENGES AND SOLUTIONS

Based on the structural components and expected functions of our predictive framework, there are at least three major challenges need to be understood properly.

Firstly, the problem of imbalanced label. The number or proportion of successful companies is inevitably small in terms of all companies in our dataset. Models trained over the dataset like that would probably learn very limited information about companies' business success, thus leading low precision or recall for the class of successful companies.

Secondly, the choice of model hyper-parameters. As we mentioned before, our machine learning system include a set of classifiers that provide their own judgement independently. So, finding out the optimal or at least a moderate choice of hyper-parameters for each model is a requirement for acceptable model performance.

Besides, another challenge will be the risk of generalization. A good performance over the training dataset does not guarantee the same results can be derived on the testing dataset. Also, a seemingly useful classifier over a relatively large set of data does not mean it could still be useful on a small set of data.

In order to deal with such problems, we try to apply a refined system design that includes the phases of oversampling, hyper-parameters tuning and robustness testing. To the best of our knowledge, such techniques can effectively detect or mitigate the negative effects from the problems we mentioned above.

V. DATA DESCRIPTION

The datasets used in this project can be roughly divided into three categories, which are company-related data, exchange rate data and web search data. The company-related data is gained from Crunchbase Data Export, the exchange rate data is collected from OANDA database and the web search data is obtained from Google Trends. The raw datasets include records of more than 20,000 companies from 68 countries, January 2014 to February 2021. To be straightforward as much as we can, here we only summarize and show the selected features as the Table I.

VI. FRAMEWORK DESIGN

Our machine learning framework can be divided into four main components. They are feature extraction, feature creation, intermediate feature processing and predictive modelling. This thinking can be summarized as the experimental design diagram shown in Fig. 1.

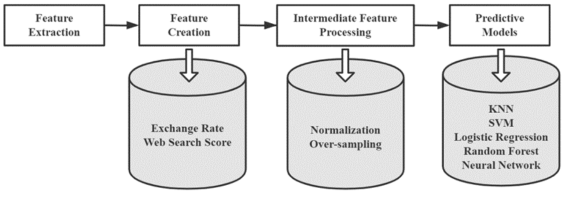

Fig. 1. Framework design diagram

A. *Feature Extraction*

Company-related datasets from Crunchbase Data Export include CSV files that related to companies' profile, investment reports, acquisition information and so forth. In order to have a good consistency between datasets, here we removed duplicates and filtered companies by restricting their founded dates between 2004 and 2015. Features were selected based on the consideration of data consistency and missing rates.

As for the exchange rate dataset, we chose gold price to be the target (basic currency) and formatted the annually data for each currency based on the country codes for each company. For the web search dataset, we formatted the monthly web search score of IPO and M&A for each country code. The time window of both exchange rate dataset and web search dataset matched the horizon of company-related datasets.

B. *Feature Creation*

When processing exchange rate data, we matched each company and respective currency by its country code, and then we further identified respective time window for calculation through the founded year of the company. Based on this, we would be able to calculate the mean and the standard deviation of the exchange rate over the five years since the company founded. Follow the same logic, we also calculated the mean and growth rate of web search scores of both IPO and M&A over 5 years since the company founded.

These calculated features not only allow us to measure the average level and the fluctuation of exchange rate in a moderate period of time since the company founded, but also provide us with search-based leading indicators of a company's business success (IPO or M&A). So, we expect such features would add value to our system.

C. *Intermediate Feature Processing*

Normalization

Due to the requirement of many classification algorithms such as KNN, SVM, NNK and so on, we need to scale our features in a proper way. For numerical features, we use z-standardization method to normalize their ranges. The normalization algorithm follows the formula:

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

In formula (1), x represents a specific value of the feature, μ is the mean value of the feature, σ is the standard deviation of the feature. By conducting this algorithm for each value, the feature will have a mean of 0 and a standard deviation of 1.

Over-sampling

Due to the issue of imbalanced label distribution, we need an effective method, over-sampling, to mitigate the problem. Fig. 2 shows the issue and one way to implement over-sampling.

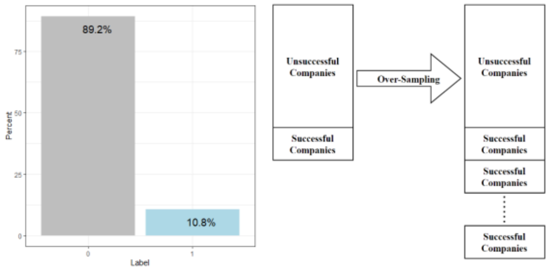

Fig. 2. The issue of imbalanced label and respective solution

This KNN-based approach multiplies the successful companies in the training dataset to form a synthetic balanced dataset. Fig. 3 shows the distribution of training labels before and after performing the over-sampling. As a result, predictive algorithms will have more opportunities to learn the business success samples in the training process over this synthetic dataset. However, one potential risk of this method worth pointing out is overfitting. So, validation part would be very important in our system.

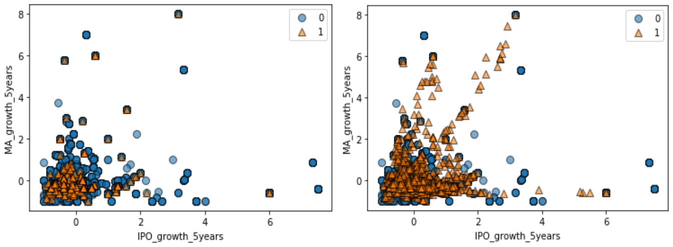

Fig. 3. Training dataset labels before and after over-sampling

D. *Predictive Models*

There are several powerful machine learning algorithms widely used in classification problems in business and finance. Here we choose five of them to build a set of competing classifiers in our system.

K-Nearest Neighbors (KNN)

KNN is a distance-based classification algorithm. Given a positive integer K and one observation x_i, the algorithm will identify the K points that are closest to x_i. Then, it estimates the conditional probability for each class and classifies this observation x_i to one class that obtain the largest probability. The most important hyper-parameter of KNN is the value of integer K.

Logistic Regression (LR)

LR formulates the binary classification problems as a regression process [5]. The simplest logic of this algorithm can be summarized into three steps. Firstly, based on the thinking of gradient descent, the model would undergo an iterative process to minimize it cost function, thus finding out a set of optimal coefficients of the prediction function. Then, through the implementation of the sigmoid function, see formula (2), LR would be able to transform each initial regression output into an estimated probability.

$$g(x) = \frac{1}{1+e^{-x}} \quad (2)$$

Finally, by setting up a specific threshold of estimated probability, the algorithm will return the predicted class.

Support-Vector Machine (SVM)

Although concerning complex computation, SVM follows a very natural intuition that separating different classes through a linear or nonlinear boundary or line. To address more common situation that nonlinear boundaries between classes, a special kernel approach will be applied to enlarger the feature space.

C could be one of the most important hyper-parameters when applying SVM. In general, when *C* is relatively small, the tolerance of violations to the margin would also be small. So, it always leads to a narrow margin. As the *C* goes up, we will tolerate more violations to the margin, thus leading to a wider margin.

Random Forest (RF)

Random forests operate by constructing a multitude of decision trees in the training process and outputting a most close class of the individual trees based on the parameters inputted. These large number of decision trees are built over bootstrapped samples[6]. If a simple decision tree model is trained on B number of bootstrap samples, then the prediction of the RF, denote as f_{RF}, will be the average of individual predictions, denote as f^*, coming from these decision trees. That is:

$$\widehat{f_{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x) \quad (3)$$

In each split, a random set of features is selected as split candidates, but only one of these predictors would be used in the split. In this way, RF algorithm has an advantage of decorrelating predictors, thus reducing the variability of the model performance.

Neural Networks (NNK)

Similar to the concept of human neurons, NNK contains many elementary units called neurons[7]. When a neuron receives signals, it will process them and change its internal state based on the signals, and then signal neurons connected to it. Signals can be real numbers, and their outputs are computed by some linear or non-linear activation functions. The connections between neurons are also known as edges. Neurons and edges have a weight to adjust the learning process. Based on this structure of organization, NNK systems will be able to learn complex patterns from data without applying any specific rules.

VII. Experimental implementation and evaluation

A. *Experimental Implementation*

As shown in Fig. 4, for each predictive model, we will firstly learn the model over the training dataset (50%), and then conduct the hyper-parameters tunning over the validation dataset (10%). Due to the cost of time and computation, here we only pick one or two parameters for each model and hope to obtain a set of moderate options after this iteration process. We do not pursue the realization of optimal performance.

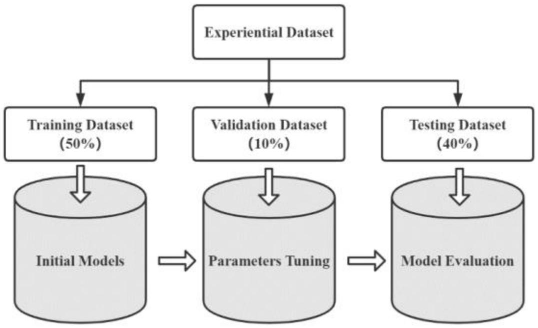

Fig. 4. Predictive modelling process

Given the fair option of hyper-parameters, we would test the out-of-sample performance for our model sets. This part will include a series of evaluation metrics and techniques. Based on the results, we will evaluate the value of some innovative features, test the robustness of our framework, and discuss the potential user scenarios for selected models.

B. *Evaluation Metrics*

Confusion matrix is a specific layout that can visualize models' performance efficiently. In general, each row of the matrix represents the instances in an actual class and each column means the instances in a predicted class. When plotting a confusion matrix, there are naturally four useful metrics come in, which are accuracy, precision, recall and F1-score.

To evaluate and compare the overall performance for different models, we would use the confusion matrix based on testing samples and mainly focus on the accuracy and weighted F1-score (a weighted combination of precision and recall). By contrast, when we dig into specific user scenarios, precision and accuracy might be more meaningful in the discussion.

VIII. Major results and analysis

A. *Validation and Test Results*

From Table II, we can see both validation performance and testing performance for each classifier trained with all features we selected. In our framework, SVM has the best performance in terms of validation and testing accuracy, KNN has the best performance in terms of validation and testing F1-score, and LR is the weakest classifier among this model set.

After performing hyper-parameters turning, most of models have a relatively moderate testing performance, and the performance difference between testing dataset and validation dataset is not large.

TABLE II
ALL FEATURES: VALIDATION AND TESTING SET PERFORMANCE

Models	Validation Performance		Testing Performance	
	Accuracy	F1-score	Accuracy	F1-score
KNN	0.8103	0.8096	0.8260	0.8359
LR	0.6753	0.7284	0.6567	0.7235
SVM	0.8696	0.8183	0.8908	0.8489
RF	0.8088	0.8209	0.8144	0.8332
NNK	0.8412	0.8341	0.8547	0.8572

B. *Robustness of Testing Results*

To further test the robustness of the framework, we also conduct a 5-fold cross-validation for each model learned before (except for time-consuming NNK) over randomly selected small, medium and large sets of the experimental data. The results are shown in Table III. We can tell that over different datasets, the accuracy and F1-score values of KNN, LR and RF are consistent, which means such models are relatively robust and have a great potential of generalization.

However, the performance of SVM shows highly visible changes over datasets with different sizes. Therefore, we can expect that the hyper-parameters setting of SVM is limited to specific datasets, thus specific model setting has weak potential of generalization.

TABLE III
ROBUSTNESS TESTING: CROSS-VALIDATION ON 1.9K, 9.5K, 19K SAMPLES

Models	10% dataset (1.9K Samples)		50% dataset (9.5K Samples)		100% dataset (19K Samples)	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
KNN	0.8959	0.8655	0.8885	0.8560	0.8832	0.8489
LR	0.8933	0.8492	0.8951	0.8486	0.8913	0.8423
SVM	0.5134	0.5527	0.1717	0.1580	0.6314	0.6135
RF	0.8820	0.8603	0.8817	0.8592	0.8770	0.8536

C. *Feature Configurations Analysis*

To evaluate whether different types of feature we obtained in the feature creation phase would add value to our system, here we also compared the testing performance for each model in our system given different combinations or configurations of input features. We firstly fed the model with purely company-related features, then we added exchange rate features and web search features successively. The results are summarized as Table IV.

TABLE IV

FEATURES CONFIGURATIONS: TESTING PERFORMANCE COMPARISON

Models	Purely Company Features		Company Features + Exchange Rate Features		Company Features + Exchange Rate Features + Web Search Features	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
KNN	0.8028	0.8154	0.8113	0.8243	0.8260	0.8359
LR	0.5099	0.5955	0.6122	0.6873	0.6567	0.7235
SVM	0.1061	0.0254	0.8947	0.8496	0.8908	0.8489
RF	0.7370	0.7788	0.8421	0.8464	0.8144	0.8332
NNK	0.8738	0.8525	0.8730	0.8555	0.8547	0.8572

Based on the initial model performance, we can tell that exchange rate features significantly improved the testing performance of KNN, LR and RF. As for NNK, the introduction of exchange rate features does not make much difference on the testing accuracy and F1 score. As we add the web search features to the system, we can note that KNN and LR show additional refinement, SVM does not display significant performance change, while the performance of RF and NNK show slightly deterioration.

Due to the occurrence of general refinement in most of the models, we can justify the value of exchange rate features in the prediction of companies' business success. Given the configuration of company-related features and exchange rate features, the effectiveness of web search features may be different in different classifiers. These data from Google Trends improve the performance of some models such as KNN and LR, but it may not always be valuable in some other classification algorithms.

D. *User Scenarios Analysis*

In addition to the evaluation of overall model performance, we also try to explore specific user scenarios of our machine learning system, thus making this application closer to the decision making in business and more useful.

Generating correct judgement on companies' business future in a reliable way is crucial to institutional investors and angel investors. So, if they plan to introduce our predictive system as a solution to support their investment decision making, the reliability of prediction would be highly important. Therefore, apart from metrics of overall performance such as accuracy and F1-score, precision is another evaluation metric we need to focus on when analyzing the predicted outcomes.

Take two typical classifiers in our framework, LR and NNK, as an example, we plot their testing sample confusion matrix as Fig. 5. As we can see, for class A (successful companies), NKK has a higher precision of 31.57%. Given the great potential of generalization, when our model returns a positive prediction, we would be able to use this reliable probability in the calculation of the expected return of investment.

Besides, we can see both models have an excellent precision for class B (unsuccessful companies), which are 95.19% and 92.29% respectively. Under the assumption of great ability of generalization, if we receive a negative prediction from each model and they make the prediction independently, we will have strong confidence about its tiny possibility of IPO or M&A.

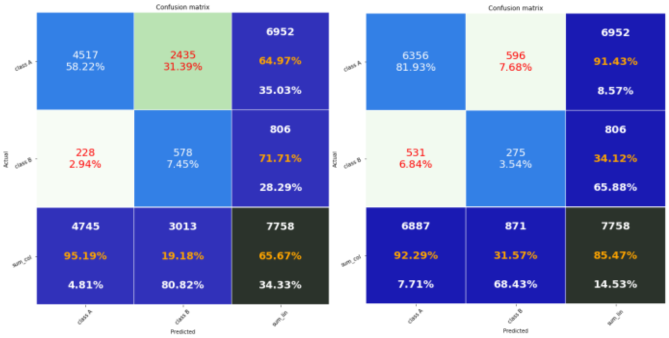

Fig. 5. Confusion matrix for LR and NKK over the testing set

IX. CONCLUSION AND FUTURE WORKS

There are many imperfect solutions thus exciting opportunities in the business applications of machine learning system nowadays. Our framework is an attempt at the refinement of ML-based business prediction and an exploration of classification solutions over imbalanced labels. In this project, we proposed a machine learning framework which combines exchange rate features, web search features and traditional company-related features together to predict companies' business success.

We firstly set up the criterium of labeling. Companies experienced IPO or M&A in the historic data would be identified as successful ones. Besides, based on our experimental design, we completed feature engineering and conducted predictive modeling and evaluation. We found that all classifiers in the framework display robustness under the experimental setting of one validation dataset, but in a different design of cross-validation, the performance of SVM was significant unstable between different datasets, which indicates its limitation on generalization.

Moreover, we evaluate the impact of feature configuration on the performance of our model set. We found that the exchange rate features add great value to all five models in terms of accuracy and F1-score, while the web search features only refine KNN and LR.

Finally, we summarize the potential user scenarios of our machine learning framework. In the case of LR and NNK, we analyzed how machine learning could assist investment research and investor decision making through reliable prediction of positive and negative class. We believe that the success of such applications highly depends on the ability of model generalization.

In the future, there are at least two aspects worth exploring. Firstly, more powerful features. Big data technologies have unleashed the great power of feature creation. Many novel features that contain valuable information remain to be constructed. Besides, the impact of time horizons. Although most of the models in our system display good robustness, it is still not clear whether their performance limits to specific choices of time window. More specialized research in this aspect would be helpful to deepen our understanding of model adaptability and robustness in broader user scenarios.

VII. Acknowledgment

We would like to say thanks to those who had done previous research in the field of business success prediction. This project was inspired in part by their contributions. Also, we want to thank all Econ 425 teaching staffs, including Prof. Xiaobai Liu and TAs for their help and assistant.

References

[1] Chenchen Pan, Yuan Gao, Yuzi Luo. *Machine Learning Prediction of Companies' Business Success.* http://cs229.stanford.edu/proj2018/report/88.pdf

[2] Francisco Ramadas da Silva Ribeiro Bento. *Predicting start-up success with machine learning*. PhD thesis, 2018.

[3] Ibukun Afolabi, T. Cordelia Ifunaya, Funmilayo G. Ojo, Chinonye Moses. *A Model for Business Success Prediction using Machine Learning Algorithms*. IOP Conf. Series: Journal of Physics: Conf. Series 1299 (2019) 012050 IOP Publishing, DOI: 10.1088/1742-6596/1299/1/012050

[4] D. Asir Antony Gnana Singh, E. Jebamalar Leavline, S. Muthukrishnan and R. Yuvaraj. *Machine Learning based Business Forecasting*. I.J. Information Engineering and Electronic Business, 2018, 6, 40-51 Published Online November 2018 in MECS, DOI: 10.5815/ijieeb.2018.06.05

[5] Ali, Kartal. Balance Scorecard Application to Predict Business Success with Logistic Regression. Journal of Advances in Economics and Finance, 2/1/2018, Vol.3 (1)

[6] Jhaveri, Siddharth , Khedkar, Ishan , Kantharia, Yash , Jaswal, Shree. *Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns.* 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), March 2019, pp.1170-1173

[7] Costantino, Francesco, Di Gravio, Giulio, Nonino, Fabio. Project selection in project portfolio management: An artificial neural network model based on critical success factors. International Journal of Project Management, Nov 2015, Vol.33(8), p.1744

Yiheng An received his bachelor's degree from Nanjing Agricultural University and a graduate certificate from UCLA Extension. His research interests include data analytics and business intelligence. He has done several business projects in these fields for companies such as Facebook, MediaAlpha and so forth.

Xinyi Zhang received her bachelor's degree from University of California, Irvine. She participated in a research group as an undergraduate which focused on causal inference of retail companies' success in United States during the Great Depression. In graduate school, her research direction related to machine learning, data analytics and prediction.

Beier Cao received his bachelor's degree from Ohio State University. His research interests include machine learning, data mining, financial risk modelling and predictive analytics. He has done other research projects in these fields such as using historical data to predict the future development of traditional taxi in New York City.