# FINANCIAL ENGINEERING

## *Overall Insights into Machine Learning in Credit Risk Management*

Yixin Jin, MSSP + DA ‘21

## INTRODUCTION

After training the data of loan applications and the application outcomes from 81,103 clients on LendingClub.com, a decision tree model for loan request outcomes at the accuracy level of 92.86% has been delivered. To improve profitability, other than mitigate the operation risks, LendingClub can even apply this prediction model to profile the ideal accepted customers and thus target them to expanse their businesses.

LendingClub’s data has opened the door to the application of machine learning in finance. And this report will offer you overall insights into the current state of the application in credit risk modeling, covering the commonly used features and labels, available training datasets, unsolved issues, and potential solutions.

# FEATURES

Artificial Intelligence has become integrated into our daily routine life, and recent years have witnessed how machine learning has reshaped the finance industry. With the high volume of data generated inside and outside the industry, the machine learning algorithm has been efficiently adopted in credit risk modeling. Therefore, the exploration of the credit risks modeling will begin with the available features.

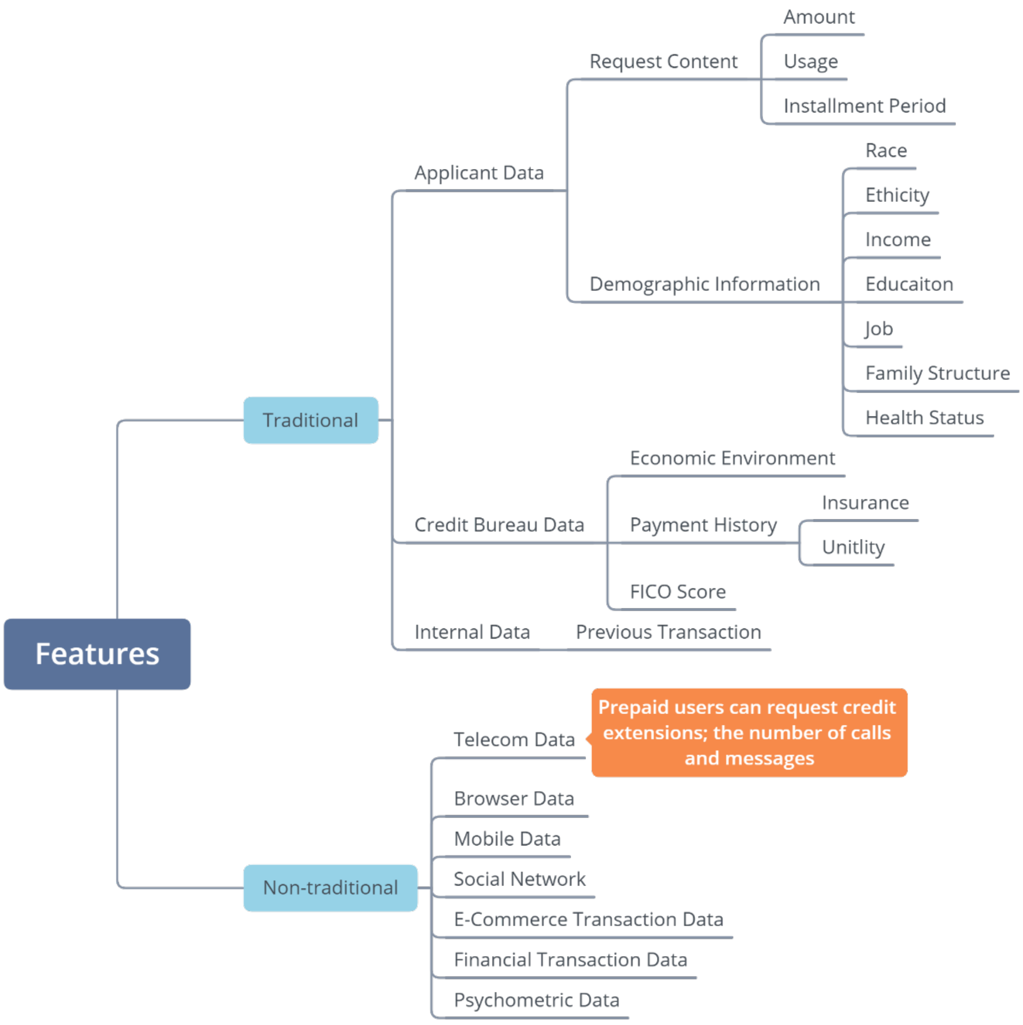

Figure 1. Traditional Features and Non-traditional Features

We can extract thousands of features from various datasets and generally categorize the features into two groups. As Figure 1 shows, one is the group of traditional data; the other is the non-traditional data. The features of the traditional group have been commonly used. For example, the data offered by the LendingClub can be a good case of traditional data. For an applicants, no matter an individual or a company, the traditional features can be applicants' past history of borrowing and repayments, or insurance results, etc.; more specific, for a company, the accounting variables or market-based variables, including market equity value, and the soft facts of the firm's competitive position or management records can be incorporated into the group of traditional features. The LendingClub dataset of 11 variables we have used in assignments has served as a good case of traditional features, which is a subset of the original dataset consisting of 56 variables (*Data Cleaning and Preparation for Machine Learning – Dataquest*, n.d.). Covering the features of demographic information and financial history information, the prediction model may perform better than my decision tree model.

However, the alternative data in the non-traditional group has been growingly incorporated into the credit evaluation system. It's reported that Experts estimate that about 45 million U.S. consumers lack the credit history required to produce trustworthy credit scores under the current system.

As an international student, I must be one member of the cohort with a limited credit history. Thanks to many companies like Lenddo, who are trying to use the information gleaned from social media, or even the writing styles as the prediction features, the lenders can assess consumers who cannot otherwise obtain credit in the mainstream credit system, thus significantly promoting financial inclusion (*Lenddo: Using data analytics to boost financial inclusion – Digital Innovation and Transformation*, n.d.). Also, with the machine learning model, the financial institutions can efficiently link personal information to the identity of a client among the universe of similar characteristics.

# TARGETS

The financial machine learning model has been applied to many credit risk management scenarios, significantly improving decision efficiency.

## Fraud Detection

Based on the data of each action a cardholder takes, the bank can assess if an attempted activity is characteristic of that particular user. Therefore, fraudulent behaviors are under monitor with high precision. In this case, the transaction can be labeled as a regular transaction or an abnormal transaction (*How Machine Learning Facilitates Fraud Detection?*, n.d.).

## Credit Scoring

Credit scoring is a target variable in the numerical format. Under the machine learning algorithm, credit scoring can be adjusted dynamically (Wang et al., 2011).

## Credit Limit Adjustment

If a client requests to improve the credit limit, the target variable of the algorithm would be whether the clients are qualified for higher credit lines.

## Loan Application

In cases like the LendingClub dataset consisting of the client information for a loan, the target variable will be whether the application can be approved or not (*Data Cleaning and Preparation for Machine Learning – Dataquest*, n.d.).

# CASE STUDY ON A REPAYMENT PREDICTION MODEL

Holding a similar belief of realizing financial inclusion as LendingClub, Home Credit, an international non-bank financial institution, is running the businesses of offering financial services to the population with insufficient or non-existent credit histories. It once hosted a data science competition on Kaggle, hoping to deliver a reliable repayment ability prediction model (*Home Credit Default Risk*, n.d.). Based on the 10 datasets covering the information of past credit histories, etc., as Figure 2 shows, many data scientists have successfully built fascinating models to predict whether a loan will be repaid. The 1st Place Winner has applied the methods of feature engineering and random to dimensionality reduction, significantly improving the efficiency of the prediction model. Also, the ROC AUC was taken as the model performance metrics (*Start Here: A Gentle Introduction | Kaggle*, n.d.). We can refer to Kaggle for more similar cases to gain an in-depth understanding of the model.

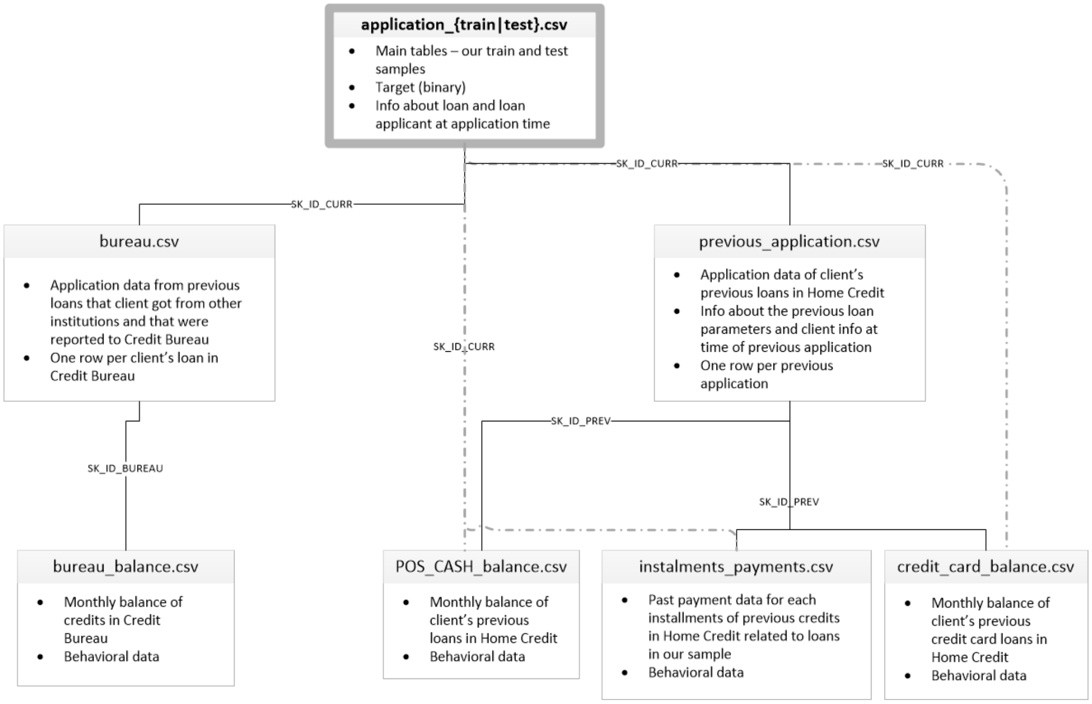
application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Figure 2. Home Credit Datasets

# RISK AND CHALLENGE

Despite the overwhelming advantages of the application of machine learning in financial credit risk management, its accompanying risks and challenges need paying ample attention to. The insights into the risks and challenges enable us to apply the models better. Such unsolved issues can be categorized into two types: public concerns and the problems that the companies are facing.

## Public Issues

### *Privacy Violation*

Under the circumstances that many companies have cooperated with each other to construct the consumer data-sharing network, customers may be unaware that the data they agree to share with developers on this platform will be further shared with other platforms (*Tencent's WeBank applying "federated learning" in A.I. - DigFin*, n.d.). If the scale of the approved permission for the companies to use the data is unclear, customers' privacy may have been violated in the data-sharing network.

### *Model Bias*

The machine learning model is constructed based on human interactive data, indicating that Take the feature of races for instance. According to a statistics report, 10.1% of Asian applicants were denied a conventional loan. By comparison, just 7.9 % of white applicants were rejected (*Fair AI*, 2019), demonstrating the loan denial rates for some ethnic groups are far higher than the average denial rate of 9.6 percent. If the data of human-based racial differences are incorporated into the training dataset, the pre-existing bias would unintentionally replicate.

## Cooperate Issues

### *Low accuracies*

In the business scenarios that require a large amount of investment or involve a large amount of target audience, the wrong prediction can bring financial burdens to the companies. For example, for a company with a leading market share of 30 million clients, a single percentage point decrease in the prediction accuracy can be equivalent to 3 wrong labels for 3 million clients, even more than the population in some states. Therefore, the model application in some scenarios requires higher accuracy.

### *Data Quality Assurance Difficulty*

With more data incorporated, the data structure will be stored in a less structured format than before (*Artificial Intelligence in Finance*, 2019). Therefore, it will require more tools to ensure data integrity and appropriateness. Generally speaking, in the case where the records with missing values account for a low proportion of the whole dataset, removing the records without compete information can serve as an efficient solution.

### *Limited access to data*

In most cases, a high number of training records serve as a solid foundation high-quality model. However, for small-scale financial institutions, the available internal information is limited due to low market share. With external data through cooperation, the quality of the prediction model can be significantly improved.

### *Talent Need*

The growing algorithm application requires more financial engineering talents than before, who share a similar educational background with the students in Social Policy + Data Analytics Cohort, not only proficient in the skillsets of machine learning, data management, and data analysis but also have domain knowledge in finance (*Talent shortage now the top* *corporate concern*, n.d.). More importantly, sophisticated talents can bridge the team of financial roles and the engineering team. Since the machine learning algorithm is a ‘black box’ for professional finance personnel, the interdisciplinary talents can help the staff interpret the algorithm outcomes, making the prediction process transparent to the clients. In this case, the declined customers can find the right direction while change their behavior to improve the credit scores.

# CONCLUSION

The preliminary insights can offer policymakers and finance industry some practical implications. For government institutions, more policies focusing on technology protection and privacy protection should be put forward to develop a stable environment for the emerging technology development of credit risk modeling. For the industry practitioners, they should embrace the cutting-edge applications positively through improving data quality and model performance, restructuring the talent system, and strengthen data sharing cooperation.

# REFERENCE

*Artificial Intelligence in Finance: Five Opportunities to Take the Leap | Accenture*. (2019, September 4). Accenture Finance & Risk Blogs. https://financeandriskblog.accenture.com/finance-accounting/artificial-intelligence-in-finance-five-opportunities-to-take-the-leap

*Data Cleaning and Preparation for Machine Learning – Dataquest*. (n.d.). Retrieved February 13, 2020, from https://www.dataquest.io/blog/machine-learning-preparing-data/

*Fair AI: How to Detect and Remove Bias from Financial Services AI Models*. (2019, September 11). Finextra Research. https://www.finextra.com/blogposting/17864/fair-ai-how-to-detect-and-remove-bias-from-financial-services-ai-models

*Home Credit Default Risk*. (n.d.). Retrieved February 13, 2020, from https://kaggle.com/c/home-credit-default-risk

*How Machine Learning Facilitates Fraud Detection?* (n.d.). Retrieved February 13, 2020, from https://marutitech.com/machine-learning-fraud-detection/

*Lenddo: Using data analytics to boost financial inclusion – Digital Innovation and Transformation*. (n.d.). Retrieved February 13, 2020, from https://digital.hbs.edu/platform-digit/submission/lenddo-using-data-analytics-to-boost-financial-inclusion/

*Start Here: A Gentle Introduction | Kaggle*. (n.d.). Retrieved February 13, 2020, from https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

*Talent shortage now the top corporate concern*. (n.d.). Retrieved February 13, 2020, from https://www.computerweekly.com/microscope/news/252456097/Talent-shortage-now-the-top-corporate-concern

*Tencent's WeBank applying "federated learning" in A.I. - DigFin*. (n.d.). Retrieved February 13, 2020, from https://www.digfingroup.com/webank-clustar/

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. https://doi.org/10.1016/j.eswa.2010.06.048