University of Nottingham Ningbo China

Faculty of Business

Academic Year 2019/20 Semester

Machine Learning and Predictive Analytics

(BUSI4391 UNNC) (SPC1 19-20))

Lecturer: Dr Xi Chen

Students: Mingyuan Zhu

20219292

1、Executive Summary for Continuous Risk Evaluation of Loans

Lending Club, as a mainstream peer to peer market leader and financial investment bridge, provides unsecured personal loans for borrowers and sufficient interest for investors. The decision process of loan origination is given to the private lenders and borrowers, and portals such as Lending Club. For entrances, they aimed at matching the lowest available cost of capital with the right borrower and attracts investors. Although borrower information such as borrower's personal information and their loan purpose is transparently available on risk control, identifying the benign users is still a tough question for its small loan amount, large crowd, and short period cycle. This characteristic leads the severe recognition for highest risk segment for distinguishing users, and its behaviour estimates among them.

To show the strength, growth potential and opportunity for expanding market, Lending Club focused on robust marketplace model enabling to generate and convert demand efficiently while managing price and credit risk effectively. Such enhanced operating leverage and capacity will make cash with efficiency initiatives.

So based on lending data before loan issued and time-series payment data from 2012 to 2013, especially for that borrower "charged-off", this analysis apply the risk evaluation prediction on their loan status and whether the issued loan will be paid next month. We finally will provide more accurate risk control prediction for subdivided people and effective way for solving wicked loan return control problems.

Analysis steps are as follows:

1. Preprocessing the user's historical behaviour data

2. Feature engineering operations on past user data

3. Perform feature selection on the sample set where the construction feature is completed

4. Divide training set data and verification set data according to historical behaviour;

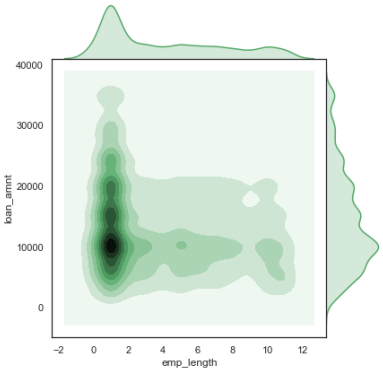
5. Establish multiple machine learning models and perform model fusion;

6. Through the established model, predict whether the user will overpay in the next month based on the user's historical behaviour data.

2、Data Exploration

For the variables which are meaningful in the business perspective, here we apply the statistics and describe summary by plotting and table. By understanding the operative side of the business, we identify the essential features that would be used for explanation for its loans status.

- For **loans amounts** of borrowers applied, committed and actually got, they have similar values for each row. The overlap distribution and count show those features have a similar distribution, which shows the money borrowers' need was well satisfied.



For the right joint plot between the loan amount and employment length in years, it shows a definite pattern for its variance distribution, as the one-year loan always with lower loan money and higher variance.

For relations between **loan status** and **interest rate**, "Charged Off" loans have a higher average interest rate of 17%, while "Fully Paid" loans only with an average 14% interest rate.

- **Annual income**:

Empirically, income is the driven and limitation features for borrowers. For the statistic summary of average salary, although the variance is huge from 9000 to 825000, most applicants with small amount of income and 75% of them are less than 88000 income which is only one-tenth of the maximum.

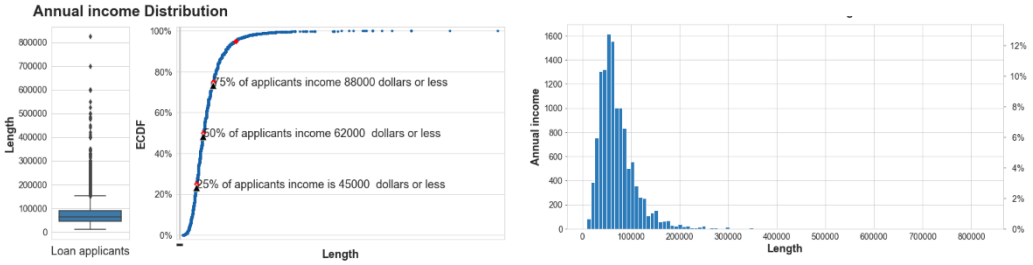


Figure 1: Annual income distribution

Interest rate: For relationship with the loan amount, the violin plot shows that borrowers that made part of the high-income category took higher loan amounts than people from low and medium income categories. Of course, people with higher annual incomes are more likely to pay loans with a higher value. The lousy loan always with a higher interest rate, as the right figures shows.

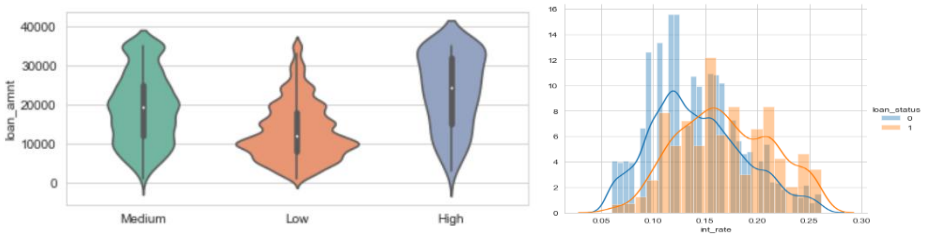
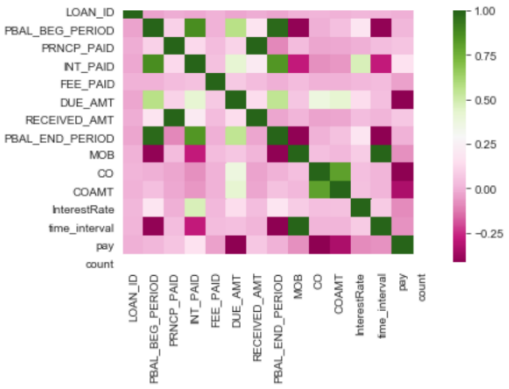


Figure 2: Violin plot of loan_amount and box plot of int_rate in different income levels

- **Time-series data:**



The right correlation heatmap shows the 2D correlation matrix of time series data. We pick the time interval related features, like "DUE_AMT", "Unreurned rate" etc. as our analysis features.

Unreturned rate: Shift index by the desired number of periods with an optional time-frequency to compare the time series with a past of itself. The lousy loan rate seems to have intricate patterns with both steady trend at the beginning and higher variance after 55. In contrast, the amount for each period performed a dropping trend with seasonality.

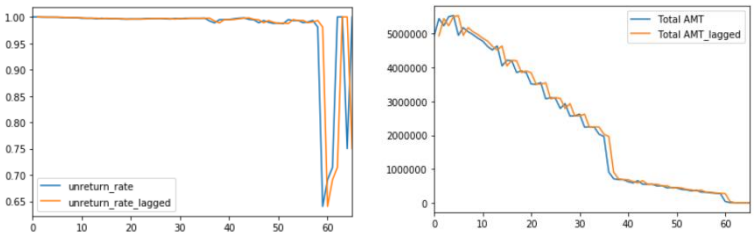


Figure 3: Shifting and lags time series of "unreturned_rate", "total_ATM"

Summary:

For loan grade, the lower class always received more substantial amounts of loans and more interest rate and even more default rate.

Its primary factors, such as a low credit grade, loans time or a high debt to income could be possible contributors in determining whether an investment is at a high risk of being defaulted.

3、Data Preprocessing

- Drop the **leak information**:

If models can reach unrealistically-high levels such as almost 100% accuracy during testing, it may imply that there is leak information hidden in our data set. Data Leakage refers to the share information between the test and training data-sets, or not divided between them. Likelike debt settlement and it indicates the compromise of the return part after default and only exists in default loans. We here drop the duplicates leak features, and generates the list in a table and "Read.ME" file.

- A multi-dimensional gap of **missing values**:

There exist several records for which certain features have no data. That is because some variables are not applicable for the specific record or introduced at a later date and hence earlier records have missing data or data was just directly not available for its record. We first drop the features which contain null values more than 900, which contains little amounts.

Then we check the null values with random distributed, and gap these by tree-based methods- "MissForest". We also gap them by propagating last, before and forward valid observation and delete features that not work by those handles. This steps we successfully drop elements from 151 to 79

- Drop the **categories features** with too many groups: A large number of classes will results in the Curse of dimensionality after one-hot. Here for the functionality is the object, we print the counts and name for each feature and drop the element which is substantial than 40. Here we lose ten features

- Data **edited**: Data editing step has compiled formats into the correct data types. For time-series data, we also create a new variable of time interval in "PMTHIST" file- the date after September 1 in 2013 as well as the rate for its default

- **One-hot:**

There are several categorical variables with many levels or unique values. This data has to be converted into numeric. The options available to process such variables are either feature hashing or hot-encoding or binning or a combination thereof. We adopt different encoding methods for different data types

  - For ordinal features, it has meaning for orders, like the grade of A, B and C, we assign ordered numbers to these values, like 1,2,3

  - For normal categories features, like the gender of man and woman, with no meaning, we just encode those categorical features as a one-hot numeric array

- **Train and test split**

This research applied "StrattifiedShuffleSplit" packages to have approximately the same ratio of "charge-off" loans compared to "fully-paid" investments in both training and testing data. For this data, it has over 84.69% of the loans considered good loans, so it is crucial to have this same ration across training and testing sets. Finally, we got "X_train" of shape (8862, 89) and (3798, 89) for the test.

4、Feature Engineering

This part, we apply the strategies of dropping the features have higher correlations and low variance and select the most import features. The top features and justifying (scores) list also provided in a notebook, including "int_rate", "subgrade", and consistent with our previous data exploration results.

- Reduce the **correlations**:

  One crucial step is identifying variables that do not qualify to be predictors. Many simpler models do not perform well, applying highly correlated variables as predictors. Scatter matrix provided for better understanding of the relationship between each other. Here based on the Pearson correlations scores, we removed highly correlated predictors and those that with only one unique value more than 0.95 ratings and print the list of drops.

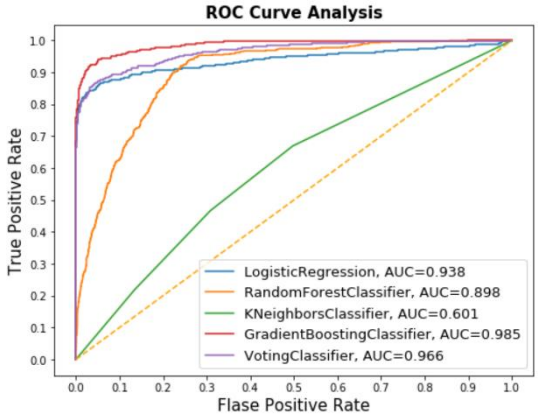
- Removing features with low **variance:** We here drop the elements with a small variation that means if a large part of features (nearly) always have the same values. We shoot three features with low variance by "VarianceThreshold".

- **GBDT**: The features which we extracted and used from the data as in input was from gradient boosted decision trees (GBDT) selections. This feature selection models performed well in extensive data set(like our loan data), and not prone to overfitting. By removing redundant features like the highly correlated and low variance in previous steps, our pipeline helps to reduce the functions set and improve the performance of the GBDT. We only keep the features with GBDT important scores more than 0.0001.

5、Model fit

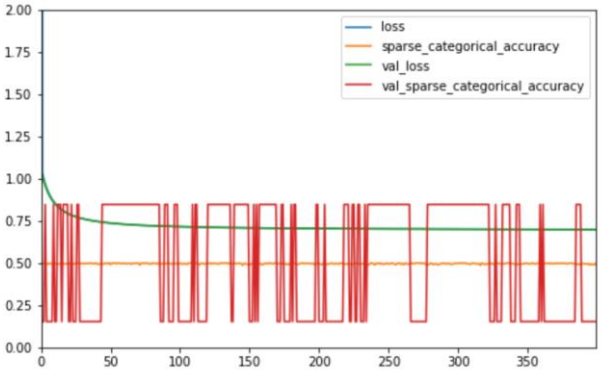
- **Oversampling:** Next Phase of analysis is to oversample our data using SMOTE Technique. We compared the test accuracy on Logistic Regression, and find that SMOTE methods to improve the performance from 94.1% to 95.7%.

- **Model fit:** Here we pick some tradition machine learning models like Random forest, KNN, SVC, Logistic Regression, Gradient Boosting, Ensembling, XGboost with bayesian-optimization, Bagging Classifier and Adaptive Boosting Classifier to evaluate whether it is useful. XGboost with auto parameters performed well with an accuracy of 0.974. The performance (accuracy) on



training as well as test set is best given using the Gradient Boosting model at 98.5%.

- For Neural Network test on different parameters, the well-performed model converged at Epoch 194 with validation sparse categorical accuracy of 0.9803.



-

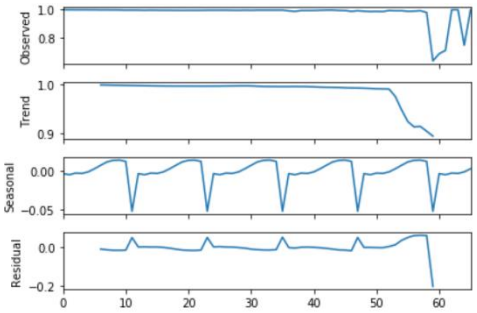
Figure 4: Training process of neuro network

6、Time-series predictive analytics

For temporal data, although they do not exist explicitly leaking information, we must take care of dependencies between your test and train set. We cannot train on future point and test on before. Otherwise, we have leaked information.

For the time-series analysis, we mainly focused on the unreturned rate, average interest rate and total due amount during each period.

For the default rate, it shows an intricate pattern among them. For applying time series decomposition and random walks using moving averages on the scale of bad loan, we can draw on conclusion as follow. These are the components of a time series



- The observed curve is steady but fluctuated after 55
- There is a steady trend at the beginning and decrease trend after 50 in the above plot.
- It's clear to see the uniform seasonal change.
- After about 55, there is non-uniform noise

  that represents outliers and missing values

To tests the null hypothesis that a unit root is present in a time series sample, here we apply the ADF(Augmented Dicky Fuller Test). As unreturned rate has p-value 0.999086, which is more than 0.05, the hypothesis rejected, and this is a random walk for its total trend.

KNN with Classic DTW: we use the models with winders of 7 days and seven days after as our predict labels, we test on the five windows, and get the results of prediction of the amount of 94.9%

VAR models: Here we apply the Vector autoregression (VAR) models to capture the linear interdependencies among multiple time series features(['time_interval', 'unreturn_rate', 'avg_rate', 'Total AMT']) among 66 observation.

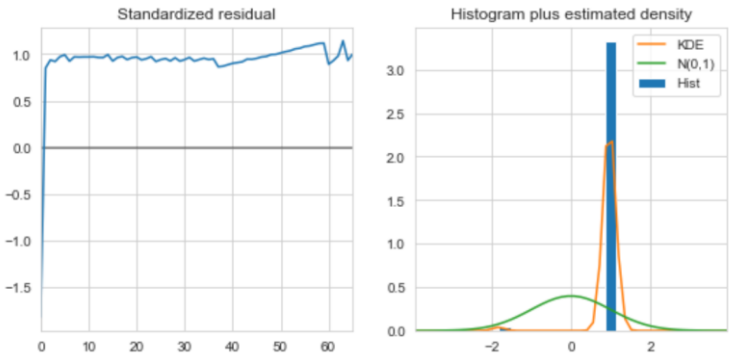
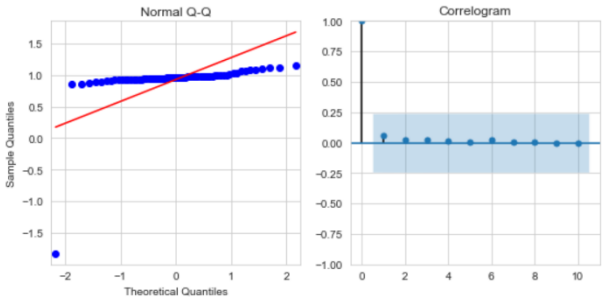


Figure 5: standardized residuals and histogram plus estimated density and Var model

A typical diagnostic for overall levels is plotting of the standardized residuals and evaluate whether they obey the model's assumptions. By testing normality absence of serial correlation absence of heteroscedasticity, it shows the two series exhibit trends or regular seasonal patterns. Therefore we assume that they are not the realization of a stationary VAR(p) process. For histogram plus estimated density, kernel density estimation(KDE) compared with N(0,1) as a reference, it shows that the distribution is more focused on 1.



The Q-Q plot, the points gather along the line in the extremities. This behaviour means our data have less extreme values compared with Normal distribution.

Correlogram indicates that autocorrelation is little likely for this set of data where an error at one point in time travels to a subsequent point in time.

7、Insights into what differentiates a loan to be risky

Credit risk refers to the possibility of a loss resulting by borrower's failure to repay a loan or meet contractual obligations, and it can be measured by credit history, capacity to repay, capital, the loan's conditions, and associated collateral. In this case, we wanted to be able to classify whether loans would remain return after a default of the loan term. We aimed at maximizing the return amount, also with more tolerance

for scoring to attract more borrowers and investments which can understand the higher risk of a bad loan. For higher risky loan, we can decrease the load time from our model performance.

For the best performance models- "Gradient Boosting", we can plot its features' rank.

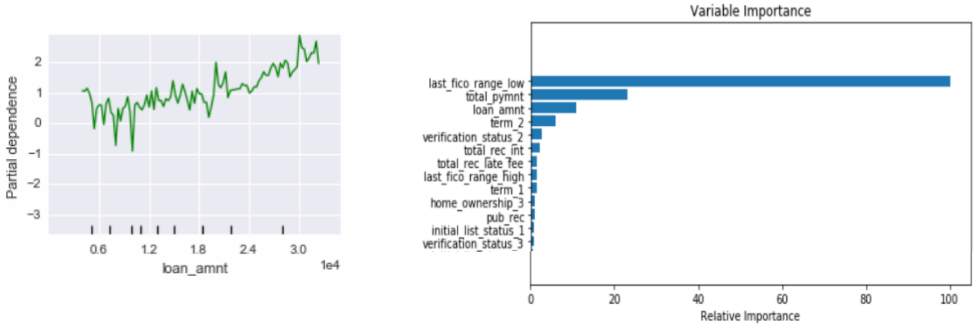


Figure6: Partial Dependence Plots and feature importance

Those features can provide good interpreter with the what decide our y- the loan

states. As expected, the most important features are those most importance variable for interpreting models including "last_fico_range_low", "total_pyment", "loan_amnt" and "term". Surprisingly the "sub-grade" does not seem to figure in the list of top 10 variables of importance. This feature is workable but not include everything. We see that there are only a handful of features with significant importance to the model, which suggests dropping many of the features without a decrease in performance. Feature importance not interpret model or perform dimensionality reduction, but model takes into account for its predictions.

Last- FICO Score variable figures as the top of the variable importance list followed by payments as one of the more critical variables in determining whether loans were eventually issued. The feature of "last_fico_range_low", that we haven't mentioned before is of vital importance for our prediction of a bad investment. It refers to the lower boundary range the borrower's last FICO pulled(score of Fair Isaac Corporation), and as the widely used consumer credit scores by payment history, accounts owed, length of credit history, new credit, and credit mix. The bank card utilization always regarded as the powerful linear effect on lending FICO scores. The creditworthiness of borrowers still has the low and medium levels bank card utilization, while higher risky borrowers always with little or over the level of card utilization with decreased funding probability and increased interest rates.

So how to reduce the risk minimization:

1 Based on Partial Dependence Plots, focus on the critical features that we selected and try to fit the data into the models. For its predictions with probability, we make a balance

2 Low risky also means they will payback in the feature instead of never paid. We can fit their time-series pattern in our model, and evaluate their performance

3 Based on that, we can provide simulator and its score for our loan