Case Study: Predicting clients loan repayment ability

(Kaggle problem statement link)

Overview

- This problem statement of predicting clients loan repayment ability is a Kaggle competition happened in year 2018. This competition was conducted by Home credit group. Home credit is an international consumer finance provider company founded in 1997 with operations in 9 countries. Home credit focus on responsible lending, primarily to people with little or no credit history. Home credit offer point-of-sale (POS) loans, cash loans and revolving loan products. Home credits aim is to provide innovative retail financial services with a focus on mass-retail lending and encouraging economic development through supporting domestic consumption, thereby improving living standards.

- Many people struggle to get loans due to insufficient or non-existent credit histories. Unfortunately, this population is often taken advantage of by untrustworthy lenders. Home credit accesses repayment ability of this unbanked population by using variety of data including telco and transactional information. Doing so will ensure that clients capable of repayment are not rejected and that loans are given.

- In this case study we will use the data provided by home credit to predict client's repayment capability. This data is as follows:

File Name	Description	Number of features
Application_train.csv	Information about loan and loan applicant when they submit the application	121
Bureau.csv	Application data from previous loans that client got from other institutions reported to Credit Bureau	17
bureau balance.csv	Monthly balance of credits in Credit Bureau	3
previous application.csv	Information about the previous loan and client information at previous time	37
POS CASH balance.csv	Monthly balance of client's previous loans in Home Credit	8
instalments payments.csv	Previous payment data related to loans	8
credit card balance.csv	Monthly balance of client's previous credit card loans	23

- Business Constraints:

1. No strict latency requirements
2. Prediction probability is important
3. Results interpretability is important

- Performance metrics:

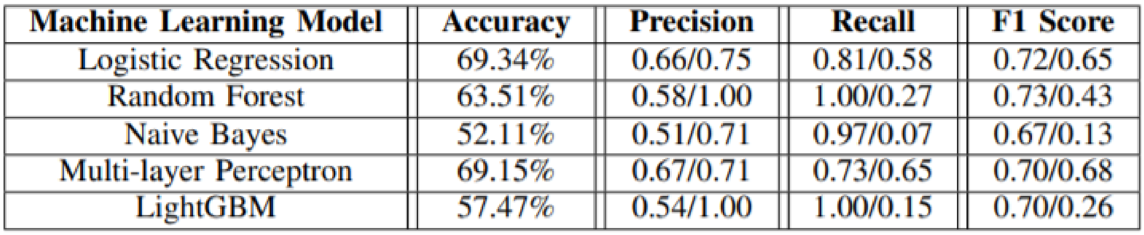
1. Area under the ROC curve
2. F1 score
3. Confusion matrix

Research-Papers/Solutions/Architectures/Kernels

1. Liang, Yiyun. “Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods.” (2019).Research Paper Link

Observations:

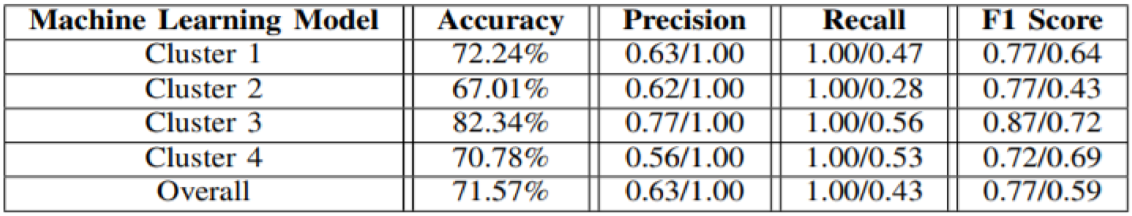
a. In this paper, initially data preprocessing techniques such as feature encoding and normalization, Invalid/empty entry replacement, polynomial featurization are carried out. Different sampling techniques are carried out and performance of various machine learning algorithms are checked. It is observed that down sampling technique worked well for this problem. Table 1 lists the performance of different machine learning models and shows that, logistic regression gives a good accuracy followed by MLP and random forest.



Machine Learning Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	69.34%	0.66/0.75	0.81/0.58	0.72/0.65
Random Forest	63.51%	0.58/1.00	1.00/0.27	0.73/0.43
Naive Bayes	52.11%	0.51/0.71	0.97/0.07	0.67/0.13
Multi-layer Perceptron	69.15%	0.67/0.71	0.73/0.65	0.70/0.68
LightGBM	57.47%	0.54/1.00	1.00/0.15	0.70/0.26

Table 1. Performance of different machine learning models

b. In this paper, K means clustering is tried to understand the machine learning models performance on different groups. Different clusters of data are formed using K-means clustering and then LightGBM model is used on these clusters to estimate models performance as mentioned in table 2.



Machine Learning Model	Accuracy	Precision	Recall	F1 Score
Cluster 1	72.24%	0.63/1.00	1.00/0.47	0.77/0.64
Cluster 2	67.01%	0.62/1.00	1.00/0.28	0.77/0.43
Cluster 3	82.34%	0.77/1.00	1.00/0.56	0.87/0.72
Cluster 4	70.78%	0.56/1.00	1.00/0.53	0.72/0.69
Overall	71.57%	0.63/1.00	1.00/0.43	0.77/0.59

Table 2. LightGBM models performance on different set of clusters

c. For different models, ROC and AUC are shown in fig. 1.

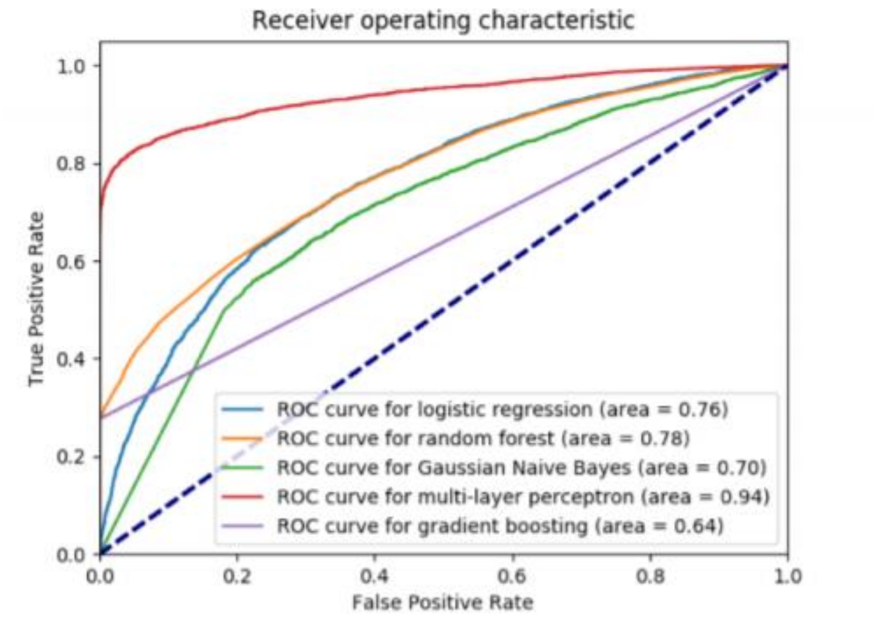


Figure 1. ROC and AUC for different models

Takeaways:

I. As the data is imbalanced, different sampling techniques needs to be used to get a satisfactory performance from the machine learning algorithm.

II. An innovative approach of clustering the data based on K-means and then using machine learning algorithm on different clusters for classification can be used to generate the better models.

2. First place solution (Link for first place solution)

Observations:

a. Team achieved 0.8057 AUC score

b. Team mentioned that there are two important things to build good machine learning model for this competition: 1. Good set of smart features, 2. Diverse set of base algorithms.

c. Ryan created multiple features based on aggregate, division, and subtraction of provided data. Label encoding for categorical feature helped Ryan.

d. For Oliver, one of the highest scoring features for the model is computed yearly interest rate.

e. For Phil, most influencing feature was the mean target value of 500 closest neighbors for each row, where neighborhood was defined by other features.

f. With the featurization technique, thousands of features were generated and many of them were redundant, or noisy or both. Forward feature selection techniques are used to do feature selection.

g. Base models are developed using stratifiedkfold with 5 folds. Models developed included classical machine learning model as well as neural network models. These around 90+ base learners were ensembled to get the final predictions. The final prediction was an equal weighted blend of ensembled 3-layer predictions.

Takeaways:

I. New features should be created using the aggregation, subtraction, and division of the provided data to improve the CV score.

II. One of the important features is yearly interest rate computation.

3. Third place solution (Link for third place solution)

Observations:

a. Team achieved 0.80511 AUC score

b. Evgeny developed a separate model for each block of database. Ensembled model consisted of these sub models developed on each block of database and a single model developed on bunch of features to do predictions.

c. Total number of features used by Evgeny were around 250. Out of the total features created, forward feature selection is used to get the important features.

d. Bureau and previous applications are the table which have many rows for single application. Instead of aggregating the features, model is developed on each row and then predictions were aggregated for a particular application.

e. Second team member of this team, Alijs used mostly statistical (aggregate) features for the prediction model development.

f. Alijs ensembled model consisted of two stacking. In the first level, 7 uncorrelated models average score is used to get the CV score. In the second level, 4 different models (LightGBM, Random Forest, Extra Trees, Linear Regression) average score is used to get CV score. Each second level models are developed on slightly different set of model predictions from the first level and some selected raw features.

Takeaways:

I. Forward feature selection is used by team to do feature selection.

II. Models developed by Evgeny were quicker because of simplicity of developed model, which allowed faster forward feature selection.

III. Ensembled model gave a boost the CV score.

IV. Second level of ensemble can be developed on the first level model predictions and some raw features.

4. Featurization

Second Place solution features link

Featurization by Pravin Kotha medium blog

Featurization by Rishabh Rao medium blog

Observations:

All the referred literature emphasized on the featurization from the provided data. In this section we will note down the observations from three sources regarding featurization.

Application_train.csv and application_test.csv are the main files which has the details of each applicant. Other provided tables are relational tables with one-to-many relation. Following are the features created by references on the provided data:

A. Current Application data

Features mentioning whether the income is grater than the credit amount, percentage of credit to income amount, annuity income percentage, days employed of applicant life (days employed/days from birth) are created from application data.

B. Bureau data

As there are multiple previous credit applications for a single applicant in bureau data, various aggregation techniques are conducted on bureau data to create a new feature. These features include count, min, max, mean of grouped bureau data for every application. Also, these features are formed considering combinations of various time durations like (aggregation function over 1 year, 2 year and 3 years etc.). Feature mentioning last application and first application year are also captured. Number of types of loan applied by applicant previously is also captured as one of the features. Debit over credit ratios are used from the calculated aggregated feature for featurization.

C. Previous application data

Similar aggregate features are created as mentioned in bureau data.

D. Positive cash balance data

Similar aggregate features are created as mentioned in bureau data.

E. Installment payment data

This data table does not have direct relation with application table, however it has a relation with previous application data and bureau data. For this table, aggregate was performed over previous application ID and then merged with application data.

F. Credit card balance data

Similar aggregate features are created as mentioned in bureau data.

G. Other features

Once the entire data table is built using the above features, aggregate features are formed on the final data frame considering the different time frames.

Takeaways:

i. As the relational tables are one to many, left outer join is performed while joining the tables.

ii. Most of the data columns have more that 50% missing values. Multiple imputation techniques are tried by researchers like mean, median, mode, model-based imputations however it is found that simple imputation by 0 value gives the best modeling results.

iii. For relational tables aggregate features were created for multiple information in a table for single application.

iv. Various features created apart from aggregate features can be referred from this subsection and the provided reference links.

5. Categorical data encoding

Handling categorical data by Applied AI

Here's All you Need to Know About Encoding Categorical Data_ Analyticsvidhya Blob

It is important to encode the categorical data present in the provided dataset. Categorical data can be of two types viz. ordinal data and nominal data. Ordinal data has an inherent order while nominal data does not have carry any ordered information. Based on the type of categorical data, different encoding techniques can be used.

A. Label encoding or ordinal encoding

This technique is used when the categorical data is ordinal. In label encoding each label is converted into an integer value while maintaining the ordinal information. For N categorical variables, label encoding consists of N integers.

B. One hot encoding

This encoding can be used for nominal encoding. In this vocabulary variable array of categories is formed first and then value of 1 is assigned to present category variable while rest of the variables are assigned 0 value. For N categorical variables, one hot encoding will have N dimensional array.

C. Dummy encoding

This is like one hot encoding. If in one hot encoding, we represent one of the categorical variables with all zeros, then it is dummy encoding. For N categorical variables, dummy encoding will have N-1 dimensional array.

D. Response encoding

As part of this technique, we represent the probability of the data point belonging to a particular class given a category. So for a K-class classification problem, we get K new features which embed the probability of the datapoint belonging to each class based on the value of categorical data. Following is a mathematical formula to calculate the response encoding:

$$P(class = X|category = A) = \frac{P(category = A \cap class = X)}{P(category = A)}$$

Takeaways:

i. Referred solutions have used label encoding and response encoding based on the categorical data type.

ii. For large number of categorical data variables, one hot encoding and dummy encoding array will have large dimension. One way to reduce the dimension is to use, effect encoding which uses 0,1, and -1 to represent data.

Solution Approach

Following steps will be followed for solving this problem:

1. Understanding the problem statement and data provided

   First step would be to understand the problem statement, provided data and available solutions. This would give a good insight of problem statement. Work is already done on this step while working on abstract and will be done while doing the further steps.

2. Posing a problem as binary classification problem

   As we want to understand the loan repayment ability of an applicant this problem is a binary classification problem.

3. Defining the business constraints and performance metrics

   It is understood from the business that there is no strict latency requirement while the model results interpretation and probability understanding is important.

   For an imbalanced binary classification problem area under ROC curve, F1 score and confusion matrix, precision matrix and recall matrix can be used as performance metrics.

4. Exploratory data analysis

   EDA will be performed to get an insight in the provided data, correlation of data with the target labels, and understanding the important features that can be derived from data for classification. EDA will also help in data preprocessing.

5. Data processing

   EDA observation will be used to process the data.

6. Featurization

   a. Available data in text, numeric and category format will be futurized accordingly.

   b. Literature study has revealed that, featurization is an important step in solving this problem statement. New features will be developed by referring the conducted research.

7. Single data table creation from relational data

   As the provided data is relational data, futurized data tables will be joined on the current application data.

8. Dealing with data anomalies

   Extensive research on the problem statement has revealed that most of the researchers have imputed the missing values by 0 and obtained a satisfactory result over the other imputation technique. This can be used to deal with missing values.

9. Train test split

   Different sampling techniques will be tried out to perform train test split on the imbalanced data.

10. Feature selection

    Important features will be selected based on the statistical methods (Weight of evidence and Information Value) for feature selection and final feature selection using votes from multiple algorithms.

11. Base model development

    Different models will be de developed, and their individual performance will be checked.

12. Ensemble Model development

    It is evident from the winner's solution that ensemble model worked well. Ensemble model will be developed to achieve better cv score.

13. Submitting the model in the required submission format.

14. Risk Scorecard Development

    Risk scorecard will be developed from the developed models prediction probabilities using PDO calibration method. This risk scorecard will predict the model score between 300 to 900 and will assess the riskiness of transaction.

15. Reporting out results and conclusion