PREDICTING CREDIT CARD DEFAULTERS

By Taha Shahid

ABSTRACT

Machine learning techniques can be used as an important tool for financial risk management.

Data Science Career Track - Springboard

Abstract

Machine Learning is an emerging area of research that aims at extracting meaningful patterns from available data. This paper highlights the significance of classification in predicting new trends from voluminous data. Performance analysis of various algorithms' like Decision Tree, Random Forest, Logistic Regression, XGBoost, and ADABoost in predicting credit card defaulters is discoursed in this project. Dataset from the UCI machine learning repository comprising of 25 attributes and 30000 instances have been employed to analyze the performance of algorithms. Moreover, the effect of feature selection has also been identified with respect to each classification algorithm. It has been concluded from the experimental results that include new found features from the data and the original data both yield the information useful for prediction and the accuracy of XGBoost method is highest in predicting credit card defaulters.

Table of contents

Contents

1. Introduction

The data set selected is called “Default of credit card clients Data Set” and as the name suggest we predict which classification method / data mining technique that will give the best accuracy for the probability of default of credit card clients.

Our client will be banks in Taiwan because banks will be utilizing the analysis and machine learning models to better understand their clientele and will be able to make decisions on their customers which will benefit them from customers that will most probably default. The dataset is already provided by the UCI machine learning repository and will be acquired through downloading XLS format of the dataset from the website.

To analyze this problem, different machine learning algorithms such as Decision Trees, Random Forest, Regression and Boosting will be used to see which method has the best accuracy for the probability of default of credit card clients. Also, PCA (Principal component analysis) will be used to see if dimensionality can be reduced. Any or all methods learned in the machine learning algorithms will also be applied.

For the deliverables of the project, an IPYTHON notebook code will be provided along with the paper summarizing the findings as well as a PowerPoint presentation.

2. Literature review

There is much research on credit card lending, it is a widely researched subject. Many statistical methods have been applied to developing credit risk prediction, such as decision trees, random forest, and logistic regression. Advanced machine learning methods including bagging, ensemble, and boosting have also been applied. A short introduction to these techniques is provided here.

Decision Trees

The decision tree structure is composed of nodes and leaves. Each internal node defines a test on certain attribute whereas each branch represents an outcome of the test, and the leaf nodes represent classes. The root node is the top-most node in the tree. The segmentation process is generally carried out using only one explanatory variable at a time. Decision trees can result in simple classification rules and can also handle the nonlinear and interactive effects of explanatory variables. But they may depend on the observed data so a small change can affect the structure of the tree.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Logistic Regression

Logistic regression is often used in credit risk modeling and prediction in the finance and economics literature. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables. A logistic regression model produces a probabilistic formula of classification. LR has problems to deal with non-linear effects of explanatory variables.

3. Methods

3.1. Bootstrap Aggregation (Bagging)

Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithms that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART). Decision trees are sensitive to the specific data on which they are trained. If the training data is changed (e.g. a tree is trained on a subset of the training data) the resulting decision tree can be quite different and in turn the predictions can be quite different. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.

3.2. Boosting

Boosting is a machine learning meta-algorithm for primarily reducing bias, and variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones. Boosting is based on the question posed by Kearns and Valiant "Can a set of weak learners create a single strong learner?" A weak learner is defined to be a classifier that is only slightly correlated with the true classification (it can label examples better than random guessing).

In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification.

3.3. Ensemble

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

4. Dataset Description (Data Wrangling)

The default of credit cards dataset was downloaded in excel(csv) format from the University of California, Irvine. Firstly, the dataset had an extra first row for the variables that was creating problems for calling the columns of the dataset. The first row of the following dataset was removed to get a more concise and useful data frame.

Secondly, the decision variable, which is binary in the dataset i.e. if the customer is a defaulter or not had a very long name “default payment next month” and was renamed to just “default” for sake of calling it with ease.

4.1. Original Features

The dataset contains 23 features (explanatory variables). As per the data description these features include:

- Age (in years);
- Gender (where 1 = male and 2 = female);
- Marital status (where 1 = married, 2 = single, and 3 = other);
- Education (where 1 = graduate school, 2 = university, 3= high school, and 4 =other); and

- Credit limit.

The dataset also contains three sets of historical explanatory variables. Each of these three sets contain six features, one for each month from April 2005 to September 2005. These three sets are:

- Past payment status that month, where:
    - -2 = no consumption;
    - -1 = paid in full;
    - 0 = the use of revolving credit;
    - 1 = payment delay for one month;
    - 2 = payment delay for two months;
    - …;
    - 8 = payment delay for eight months;
    - 9 = payment delay for nine months and above;
- Amount billed that month; and
- Amount paid that month.

4.2. Outliers

Yes, there were outliers and they were determined using the criteria that was mentioned with the dataset. The outliers for example in education column were given a 5 or a 6 and there is no information given on what type of education it represented. So, anything other than 1,2,3,4 was given 5.

The outliers in repayment status were given -2, which didn't make sense at first but carefully looking at the repayments data of a single row, it was established that it meant that the client had no consumption of credit.

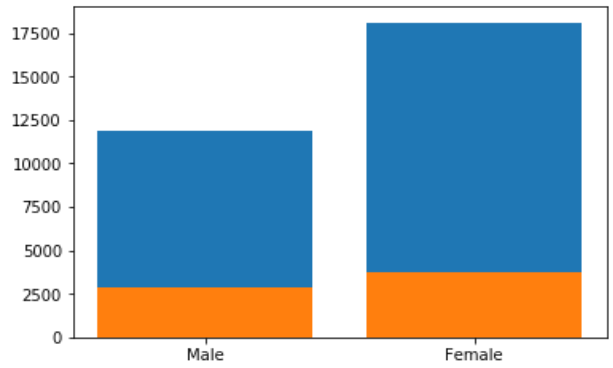
4.2. Null Values

The dataset did not contain any null values.

5. Exploratory Data Analysis (Data Story)

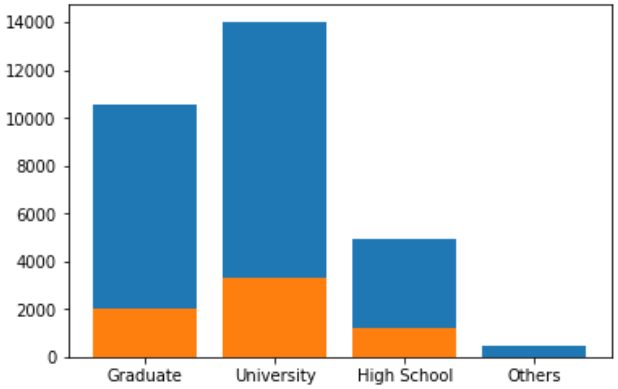
5.1. Gender Proportion

It was found out that Proportion of male defaulters is 24.16% and Proportion of female defaulters is 20.77. Also, total number of males in data was 11888 and total number of females in data was 18112.



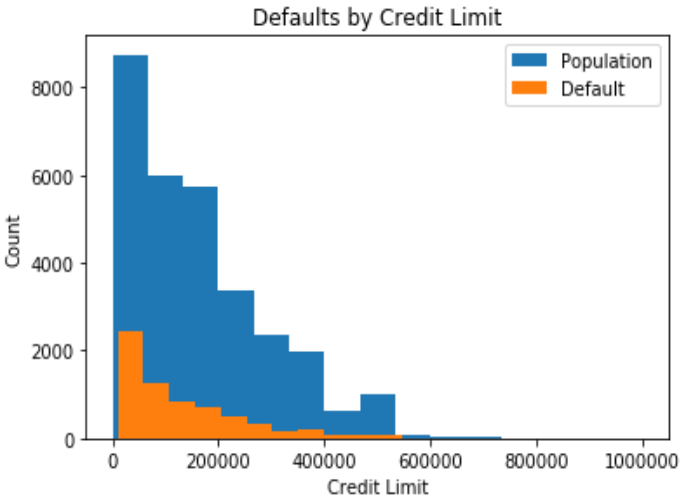
5.2. Education Level

It was found out that the highest numbers of defaulters were university students with 23.74%, second highest were High School student with 23.45%, graduate students had a defaulter proportion of 19.24% and all others has the lowest default proportion of 5.7%.



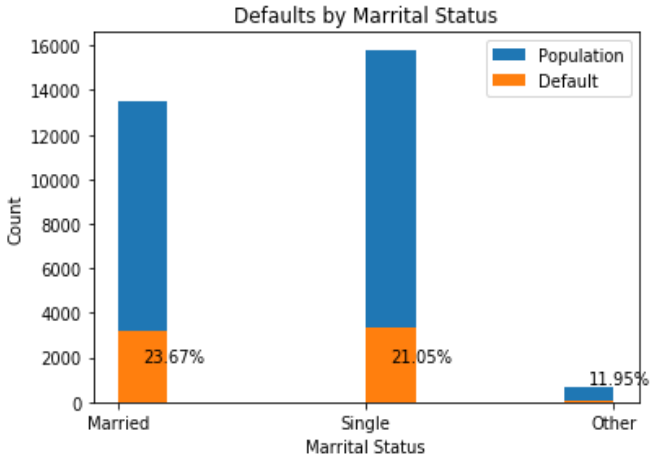
5.3. Credit Limit

Proportion of defaulters from 0 limit balance to 200000: 25.34%. Proportion of defaulters from 200000 limit balance to 400000: 15.48%. Proportion of defaulters from 400000 limit balance to 600000: 12.28%. Proportion of defaulters from 600000 limit balance to 800000: 7.69%. So, it was found out that with higher credit limit balance there were fewer number of defaulters.



5.4. Marital Status

Married defaulters were the highest at 23.67%. Single defaulters were 21.05%. And all others were 11.95%.



So, from above information it was found out that generally men had a higher percentage of default as compared to women. Secondly, students in university and high school had some of the highest percentages when it came to their accounts being defaulted. Thirdly, credit card limit balance almost had a direct correlation with the account being defaulted, as the limit balance increased lower the number of defaulters. Lastly, married and single people had a pretty close default proportion but married people for some reason had a higher proportion of default. The above finding will be investigated further using exploratory data analysis. To get a visual of the above questions investigated below are the four graphs.

6. Feature Engineering

Feature 1:

Addition of first three payments names PAY_0, PAY_2 and PAY_3.

Feature 2:

Using above three payments and making a binary column if any of the three above payments had -2 or +2.

Feature 3:

Sixth month payment minus the sixth month bill amount.

Feature 4:

First month payment minus the first month bill amount.

7. Machine Learning and Results

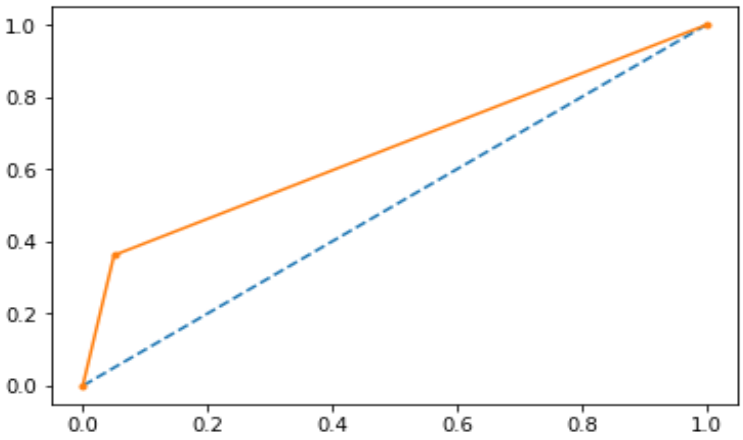
For the in-depth analysis of the default of credit card client's dataset retrieved from uci repository following machine learning models and techniques were applied.

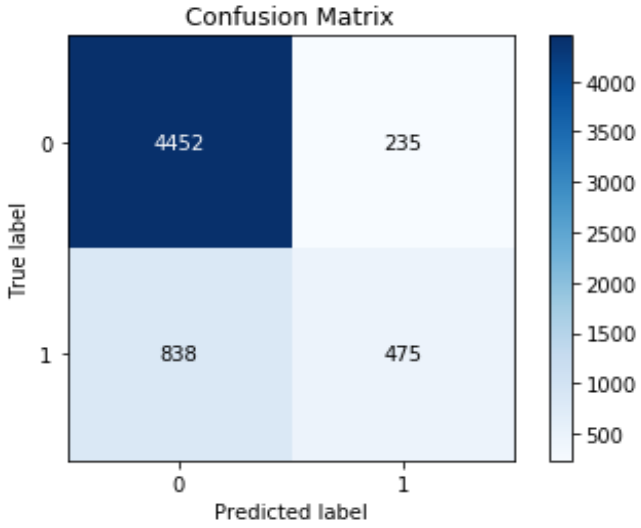
1. Decision Tree
2. Random Forest
3. XGBoost
4. ADABoost
5. Logistic Regression
6. Ensemble
7. Bagging
8. Feature Engineering
9. ADABoost after Feature engineering
10. XGboost after Feature engineering
11. Ensemble after Feature engineering

In order to fit the machine learning models, first, the dataset, which was imported as a dataframe using pandas library, was used to make X(predictors) and y(target). Second, the data was split into test and training sets using the sklearn library.

7.1. Decision Tree

Decision tree algorithm did quite well for being a simple algorithm as compared to the other algorithms performed in the analysis and it scored roc-auc score of 0.6558.





A 5-fold cross-validation score mean was 81.41%. And on the test set, decision tree accuracy achieved was 82.11%

7.2. Random Forest

Random forest ‘gini’ criterion was used with 4 jobs and 400 estimators. An accuracy of 82.08% was achieved on the test set. Also, a roc-auc score of 0.6523. A 5-fold cv score mean was 81.41%.

7.3. XGBoost

For XGboost, three parameters were optimized using RandomizedSearchCV, which is also a function of sklearn library. From optimization n_estimators were set to 57, min_child_weight was

set to 9, and max_depth was set to 3. A 5-fold cv score mean was 82.19%. And test set score of 82.02%. After feature engineering XGBoost achieved a test accuracy of 82%.

7.4. ADABoost

For ADABoost similar procedure of optimization was used as XGBoost and n_estimators was set to 162, and learning_rate of 0.001. A 5-fold cv score mean was 81.96%. Test set score was 81.93%.

7.5. Logistic Regression

After optimization, using the above-mentioned optimization in sklearn following parameters were set:

penalty='12',
dual=False,
max_iter=145,
C=2,
fit_intercept=False

A 5-fold cv score mean was 77.81%. A test set score was 78.11%.

7.6. Ensemble

For ensemble, a voting classifier was used with Logistic Regression, Random Forest, Decision Tree, XGBoost, ADABoost and parameter voting was set to hard.

Accuracy score of 81.8% was used. The best classifier was xgboost if used as is without changing any parameters.

7.7. Bagging

bag_clf = BaggingClassifier(RandomForestClassifier()
n_estimators=500
bootstrap=True
n_jobs=-1
oob_score=True
81.95% accuracy score was achieved using above classifier.

bag_clf = BaggingClassifier(DecisionTreeClassifier()
n_estimators=500
bootstrap=True
n_jobs=-1
oob_score=True
81.53% accuracy scored was achieved using above classifier.

For our analysis Decision Tree was the best classifier with mildly tuned attributes to get a 5-fold cv score mean was 81.41%. And test set score of 82.11% after feature engineering. This was the highest score achieved in this project compared to all the other standalone classifiers and ensembled method combined.

Models/Classifiers	Test Accuracy (%)
Decision Tree	82.11
Random Forest	82.08
Logistic Regression	78.11
XG-Boost	82.00
ADA-Boost	81.93
Ensemble (Voting Classifier)	81.8
Bagging	81.53

8. Conclusion

When it comes to default prediction, we have a model that can predict the defaults of customers with high enough certainty that the bank can utilize it in their functions. If the banks continue to receive customers that are represented in our dataset, we could implement our model in the banks preliminary screening process, and it would bring financial gain to the bank. However, our solution is not viable to be used as a standalone system in its current form since it only considers part of the banks actions. Many factors that were not covered in this project should be taken into consideration when taking any business action. For example, young people could be preferable for the bank since they stay longer as a customer so it could be in banks interest to favor having them as a customer even if our model would suggest otherwise. Single customers should not be discriminated against especially based on the customer segmentation which relies on calculating averages over a group. A single customer defaulting with high debt can result in much higher losses than might be anticipated simply based on averages. Similarly, the analysis does not go in-depth enough to justify if the variables used in this study could explain or predict how reliable the

customers are in the long run, especially considering that the data was collected during a debt crisis.

9. Analysis Notebooks

All of the Python code and Jupyter notebooks used in this project can be found on GitHub: https://github.com/taha-shahid/SpringBoard/tree/master/Capstone_1