STAT 5525 Course Project
A data-driven approach to predict credit card default

Hazem Sharaf
Genetics, Bioinformatics &
Computational Biology
Virginia Tech
Blacksburg, Virginia, United
States
sharaf@vt.edu

Nicolás Navarro-Navarro
Sustainable Biomaterials
Virginia Tech
Blacksburg, Virginia, United
States.
nico2710@vt.edu

KEYWORDS

Credit card default, data mining, model evaluation, credit risk

1 Introduction

Credit card usage among consumers in the United States has increased since their advent to the market in 1966 to the present, going from being a product mostly used by high income individuals to a massive adopted product used even by high risk individuals[4], and becoming most widely used method for payments in the United States [1].

Figure 1 shows the evolution in the credit card revolving consumer credit balance from 1970 to 2014, which represents the amount of credit that goes unpaid after the corresponding monthly payment [7], which also serves as indicative of the usage of credit cards over time.

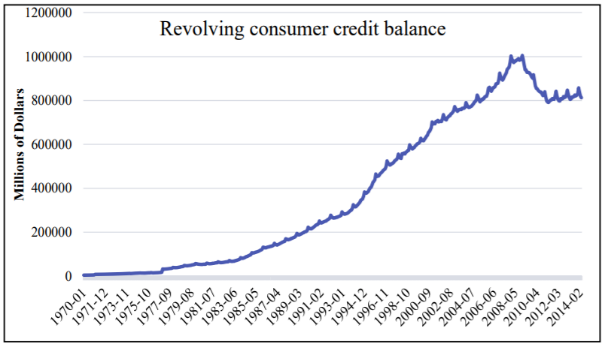

Figure 1. Revolving consumer credit balance from 1970 to 2014.

2. Response to proposal comments

The main task related to this project is classification. The classification problem consists in using attributes related to customers' personal characteristics as well as their credit card behavior in the first six months of usage to predict if the customer is going to default the following month. A credit card user defaults when no payment is made in the seventh month of usage. A series of machine learning methods provided in section 6 will be trained and tested with the data to evaluate which one among them is best adequate to make such prediction and to determine the quality of the prediction.

It is also an interest of this project to assess the difference in models' performance when the imbalance in the data is accounted for. This is a major weakness in the reference study [12], in which the word imbalance is not mentioned even once in the entire paper and issue is completely ignored.

2. Problem Statement

The purpose of the study that will be conducted is to apply data mining techniques to predict whether a new credit card holder will default on his/her credit card or not, based on several attributes and a period of usage. This will allow banks and other financial institutions to assess risk more effectively and to rely on a data driven approach for better decision making. As credit card usage has become ubiquitous in consumption payments and transactions, analyzing card holder's default becomes crucial for banks. This project will explore the relationships between different customer attributes that lead to default.

It is also an objective of this project to serve as a corroborative study to be compared to the original publication that first made use of the data set that will be utilized in this project[12].

3. Data Description

The data set selected for this study is an open access data set containing information about credit card holder attributes and performance, this is, default status. An imbalance problem is present in the data, given that there are more customers that do not default, approximately 77,8%, than those who do default, approximately 22,8%. Table 1 provides basic statistics about the data:

Attribute	Credit Line	Age	Long Term repayment
Minimum	10000	21	-708323
1st Quartile	50000	28	-20606
Median	140000	34	-1246
Mean	166123	35.45	-12715
3rd Quartile	240000	41	3142
Maximum	1000000	79	428791

Table 1. General distribution of some of the continuous predictors, including a new added attribute, long term repayment.

3.1 Attribute

The data contains 23 features, 18 of them are sextuplets of time-series: Payment codes (integer/categories), amount owed, amount paid for a preceding month. These attributes will be aggregated/derived from analysis.

3.1.1 Age: Integer, ranges from 21-79.

3.1.2 Gender: Factor: Male and female

3.1.3 Marital status: 4 levels (including 0)

3.1.4 Education level: Factor: 7 levels

3.1.5 Output Class: Binary: default or not.

3.1.6 Six additional features were added as described below in the bullet point 4.3.

4. Data Pre-Processing

4.1 Data adjustment

4.1.1 From the original paper [12], education has 4 levels, while in the dataset it has 6. Anything above 4 has been set to 4.

44.1.2 In monthly balance sextuplets, any negative balance (overpayment) is set to zero. (Affects 3932 cells).

4.2 Missing data

4.2.1 Education and marriage instances that have fields with 0, have been changed that to NA and deleted afterwards given that the building models in R were not able to handle non-available data.

4.2.2. All records where Bill amount and payment amount are zero have been removed for all sextuplets. There are about 795 instances that fall in that category. This also affected the standard deviation of the utilization rate.

4.3 Feature Engineering

The process of designing abstracted numerical fingerprints for systems of interest that reflect the underlaying structure of data based on domain knowledge is known as feature engineering in machine learning and it is one of the most important steps in the application of data analytics techniques [8]. Aiming to improve on predictive performance of the model, new attributes were decided to be abstracted for the data set using the original attributes. This will reduce number of features from 23 to 11. The following is the description of the features created for the data set.

4.3.1 Long-term payment change: It is the difference between the payment in the first month and the payment in the last month. This attribute serves as an indicator of the ability of a credit card user to pay its debts in the long term.

4.3.2 Average utilization rate: This is the average of how much of the credit limit is used per month during the six months. There is evidence suggesting that high rates of utilization credit limits can be associated with credit default risk [4].

4.3.3 Number of times where there was no need to pay: Number of months where previous balance was zero. It should be correlated to utilization rate.

4.3.4 Number of times where payment was made on time: Number of months where payment was made on time. This attribute measures the ability of a credit card customer to consistently pay its debts on time.

4.3.5 Number of times where payment was made late: The number of times where payment was submitted late in the period of six months measures how prompt customers are to pay late.

4.3.6 Standard of deviation of utilization rate: The purpose of this feature is to capture the variability in utilization rate over the six months period. The general idea is that a high variability of consumption is associated with unstable behavior, entailing greater risk.

The other variables conforming the data set are credit limit, sex, education, marriage and age. In the case of marriage, married people are believed to be more stable than single and divorced people. As far as age is concerned, early age is associate to unstable behavior and higher risk.

4.4 Data transformation

Figure 2 shows how the scale of numerical continuous variables is different across attributes. This points out the need for normalization of the continuous attributes.

The continuous attributes were normalized using the max-min normalization to scale the data into a consistent range for all the attributes and avoid distortion resulting from different scales. In addition, all category data will be presented as factor data type.

All data description and exploration are based on these modifications unless otherwise specified. In Total there are 28968 records after cleanup.

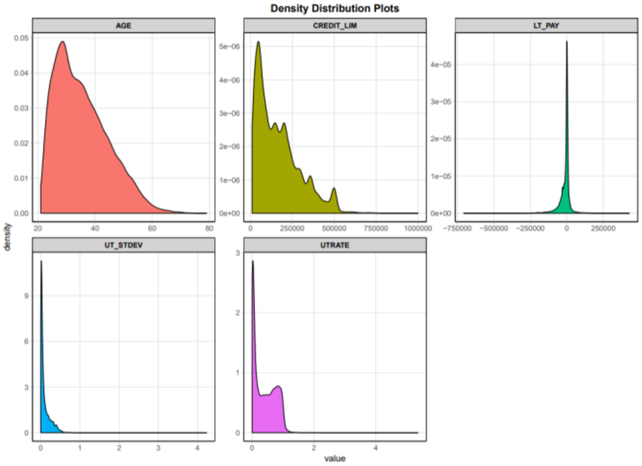

Figure 2. Distribution of the attributes age, credit limit, number of late payments, standard deviation of utilization rate and utilization rate.

5. Data Exploration

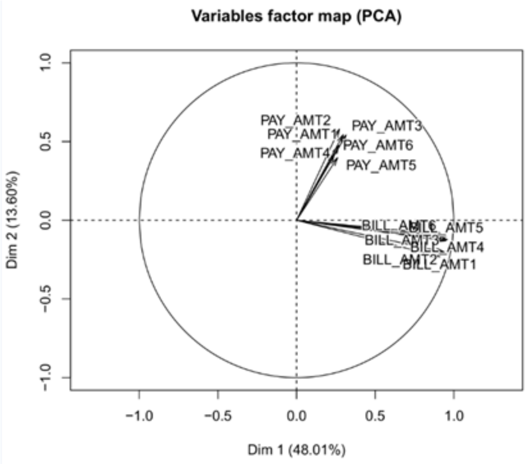

Figure 3. Principal Component Analysis of bill amounts and payments.

Principal component analysis of the bill amount and bill payments show each group of payments (sextuplets) corresponding to one of the main axes. Loading score for both sextuplets are about 0.4 on each component. They can be replaced by a single representation each.

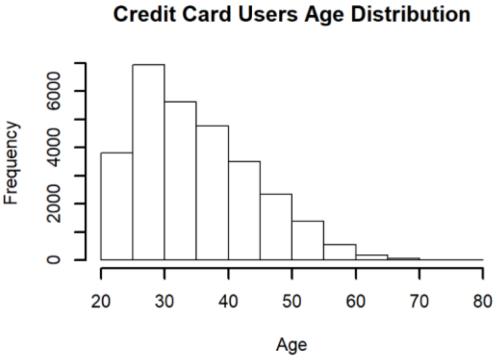

Figure 4. Credit card users age distribution.

Figure 4 shows the age distribution of the customer base for credit cards in the dataset.

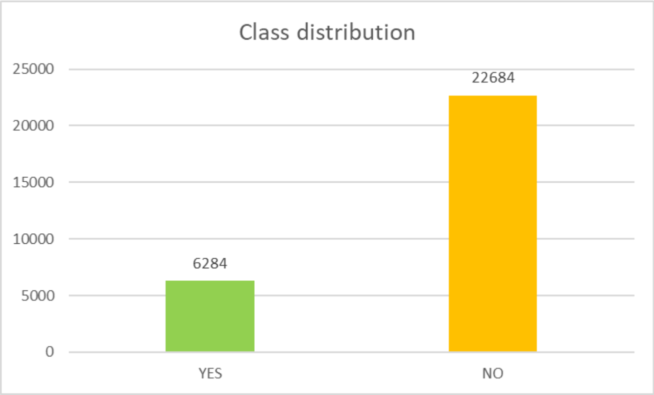

Figure 5. Class distribution in the data set.

Figure 5 shows the imbalance present in the data, where the majority class is negative, meaning that most of the credit cards customers do not default. Only 21.69 % of the customers actually default in the seventh month of usage.

6. Model Building.

The original paper referenced before in section 2 used and compared 6 algorithms to predict the possibility of default as follows. In this project, the models analyzed are decision trees, artificial neural networks, k- nearest networks, logistic regression, naïve Bayes and stacked ensemble. All the models were built using the R programming language for statistical computing.

The approach taken in this project to build the models was to try different values for the parameters that the R packages allowed to modify and see which combination provided the best results. Then, using the best performing parameters for the training phase, the models were tested using the testing set. The training results were evaluated using the accuracy measurement and the kappa value.

The training set was established as 70% of the complete data set and 30% was taken as the test set. Furthermore, a 10-fold cross validation was used for every model built to keep consistency for comparing the models. Slightly better results were achieved in the training phase for initial models with repeated cross validation, but this method was discarded given that the improvement was not significant, and it was computationally inefficient, taking very long periods of time just to train the model with one set of parameters. Results about the repeated cross validation are not shown for space constraints.

The results of the models building, and description are shown as follows:

Artificial Neural Networks: Model training			
Tuning Parameters		Performance indicators	
Size	Decay	Accuracy	Kappa
1	0.1	0.809136	0.318192
1	0.01	0.809498	0.322784
2	**0.01**	**0.809547**	**0.320745**
2	0.05	0.808659	0.315596
3	0.1	0.808512	0.311758
3	0.05	0.807229	0.305434
4	0.1	0.807131	0.304178
4	0.105	0.807377	0.306767

Table 2. Artificial neural networks model training tuning results.

Table 2 shows the tuning made to artificial neural networks. Size and decay were the parameters available to tune. In this case, the best training result was obtained when size has a value of 2 and decay has a value of 0.01. Size is the number of units that conform the hidden layer and decay is regularization parameter that limits the magnitude of the weights, making the decision boundaries in the network smoother [3].

Decision Tree: Model Training			
Tuning Parameters		Performance indicators	
CP	Split	Accuracy	Kappa
0	Information Gain	0.7632416	0.251447
0.001165038	Information Gain	0.8081661	0.3286986
0.001818595	**Information Gain**	**0.8107309**	**0.3322666**
0.003182541	Information Gain	0.8087091	0.3190981
0.0186406	Information Gain	0.8021499	0.2210436
0.05	Information Gain	0.8015584	0.1835328
0.1	Information Gain	0.7830654	0
0	GINI	0.7689113	0.2643741
0.001165038	GINI	0.8052566	0.3091185
0.001818595	GINI	0.8085114	0.3226046
0.003182541	GINI	0.8068348	0.309532
0.0186406	GINI	0.802742	0.2175584
0.05	GINI	0.8015584	0.1835328
0.1	GINI	0.7830654	0

Table 3. Decision tree training model results

Table 3 shows the tuning made to the decision tree. The complexity parameters and the split method were the two parameters available for tuning. The best performing combination was obtained when cp has a value of 0.001818595 and the method employed for splitting is information gain. The complexity parameter governs the minimum benefit that must be gained in order to split, hence limiting the size of the tree. This is a crucial step because a fully grown tree will overfit the training data, jeopardizing the testing performance [11].

Logistic Regression: Model Training			
Performance metrics	Training Model		
	Forward selection	**Backward selection**	Stepwise selection
Number of predictors	11	**10**	9
Accuracy	0.8083635	**0.8085608**	0.8084128
Kappa	0.3084232	**0.3093998**	0.3083941

Table 4. Logistic regression model training tuning results.

Table.4 shows the tuning made to the logistic regression model. In this case, there were no parameters to be modified, but three distinct regression approaches were used instead, forward selection of attributes, backward selection and stepwise selection. The best performing training results were found using the backward selection approach that uses 10 attributes.

K Nearest Neighbor Model		
Tuning Parameters	Performance indicators	
Number of K	Accuracy	Kappa
5	0.788391	0.2833777
7	0.79495	0.2874576
9	0.797909	0.2856159
11	0.799635	0.2819936
13	0.80072	0.2829218
15	0.800424	0.279188
17	0.801755	0.2815649
19	0.802298	0.2801282
21	**0.803284**	**0.2840334**
23	0.802298	0.2773392

Table 5. K nearest neighbor training tuning results

Table 5 shows the tuning made to the K nearest neighbor model. The only parameter available for tuning was the number of K that establishes the number of neighboring data points that are employed to compare a new data point and classify it in accordance. The best performing results were attained when selecting a K value of 21.

Naïve Bayes: Model Training				
Tuning Parameters			Performance indicators	
Method	Factor for Laplace correction	Adjust	Accuracy	Kappa
Gaussian	0	0	0.792782	0.36007215
Gaussian	0	1	0.792782	0.36007215
Gaussian	0	2	0.792782	0.36007215
Gaussian	1	0	0.792782	0.36007215
Gaussian	1	1	0.792782	0.36007215
Gaussian	1	2	0.792782	0.36007215
Kernel	0	1	0.782917	0.00047343
Kernel	0	2	0.783065	0.00076936
Kernel	1	1	0.782917	0.00047343
Kernel	1	2	0.783065	0.00076936

Table 6. Naïve Bayes training tuning results.

Table 6 shows the tuning results for the Naïve Bayes model. The parameters modified were the type of method utilized by the algorithm, the factor for Laplace correction and the adjustment value for kernel bandwidth. The best performing results were obtained using the gaussian method, for which the values of accuracy and kappa were the same across parameters. The Laplace correction is used in case of zero probabilities, while the kernel method is a non-parametric density estimation method compared for the normal-Gaussian estimation.

Stacked Ensemble: Model training			
Method		Performance indicators	
		Accuracy	Kappa
General linear model	No Tuning	0.808068	0.3034552
Gradient boosting machine	**Further Tuning**	**0.808906**	**0.3055586**
Random Forest	Further Tuning	0.801609	0.2982458

Table 7. Stacked ensemble training results.

Table 7 shows the training results for the stacked ensemble models. Once all the previous models were trained, a stacked ensemble method was computed having the other methods as the base learners and the ensemble as the meta-learner. Three methods were utilized for this step, a general linear regression model, random forest and gradient boosting machine. The results show that the best performing results are found when using the gradient boosting machine. Gradient boosting entailed tuning

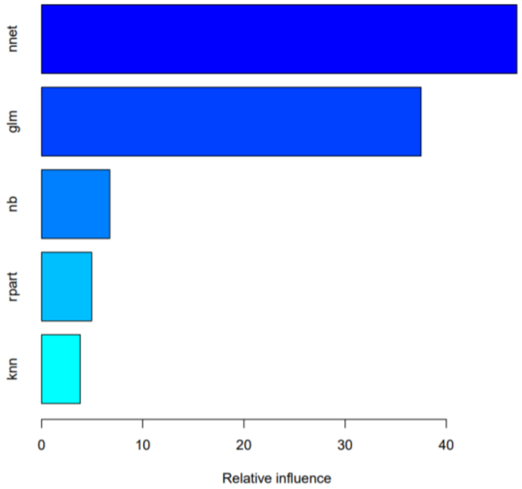

Figure 6. Contribution of each of the individual base learners to the stacked ensemble meta learners.

Figure 6 show the contribution of each of the individual base learners to the stacked ensemble meta-learner. Artificial neural network is the model that contributes the most to the stacked ensemble. It is interesting that although from the individual is Naïve Bayes as it will be shown later, it is not the one that contributes the most to the stacked ensemble.

Correlation coefficient	Decision Tree	Logistic Regression	Neural Networks	Naïve Bayes	K-nearest neighbor
Decision Tree	1	0.744131	0.797165	0.654332	0.854752
Logistic Regression		1	0.87542	0.905543	0.87375
Neural Networks			1	0.770043	0.767827
Naïve Bayes				1	0.741431
K- nearest neighbor					1

Table 8. Correlation between models.

Table 8. shows the correlation between the models that make up the ensemble meta-learner. As the pair-wise correlation coefficients suggest, there is a tendency for the models to be moderately correlated which implies that the ensemble might not provide a significant improvement in performance compared to the individual base learners.

7. Model Evaluation

The first objective of this phase is to provide an assessment regarding the relative performance of the data mining techniques to correctly predict the variable of interest, this is, to accurately reflect the underlying reality provided by the data. The second objective is to compare the outcome resulting from the formulated models to the original publication that utilized this data[12].

A crucial aspect to consider in this face is the imbalance in the data. As shown in section 3 in Figure 5, the data used in this project presents a high level of imbalance, where the majority class accounts for a total of 78.31 % and the minority class accounts por a total of 21.69 %. This situation greatly impacts the way in which the evaluation must be carried out because for example, if the data has a distribution of 90:10, where 10% of the data is the positive, then a model that predicts all the instances as negative will have an accuracy of 0.9, even though it fails to reflect that none of the minority samples were identified. This means in essence that the accuracy metric is unable to deliver adequate information about the classifier's performance given the type of classification needed as a result of the imbalanced data [5].

In order to attain this goal, traditional metrics to evaluate models will be employed. The following is a summary of the methods selected for evaluating the data mining models described in section 4:

Receiver Operating Characteristic Curve: This curve plots the sensitivity as a function of the specificity, which is the ratio of correctly predicted positive instances to the total positive instances as a function of the ratio of correctly predicted negative instances to the total negative instances. The area under the curve is then taken as a metric for the classifier performance [2].

Sensitivity: It is the probability of correctly identifying the positive samples. In this case the probability of correctly predicting that a customer is going to default.

Specificity: It is the probability of correctly identifying the negative samples. In this case, it is the probability of correctly predicting that a customer is not going to default.

Kappa: The kappa statistic is a relative measurement indicator that shows the improvement that the model represents over a random guess.

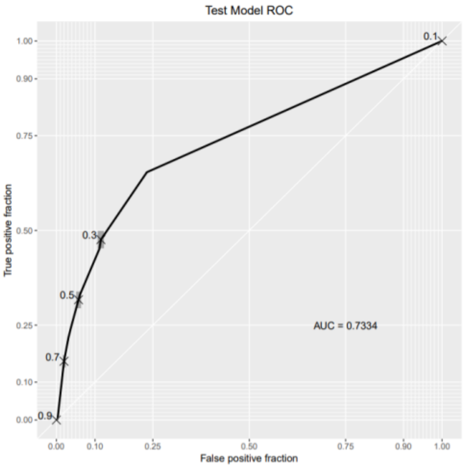

Figure 7. ROC curve for decision tree model.

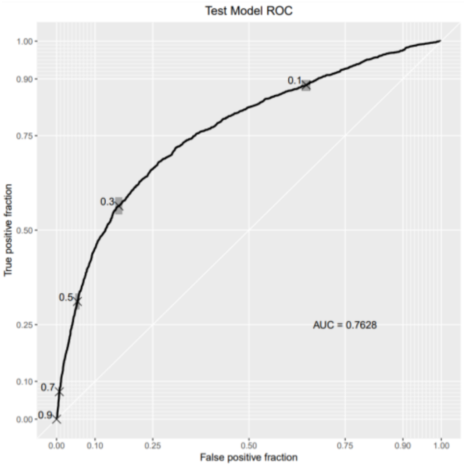

Figure 8. ROC curve for artificial neural network model.

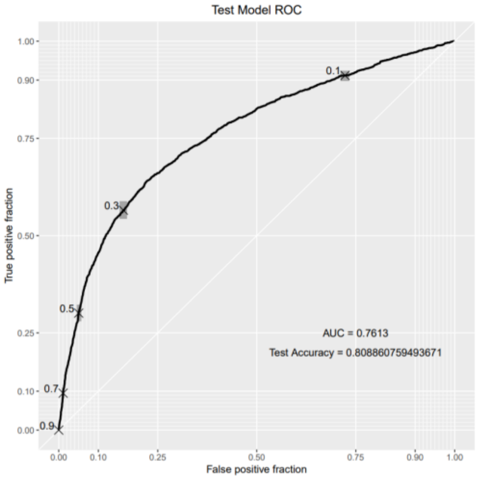

Figure 9. ROC curve for logistic regression, backwards approach.

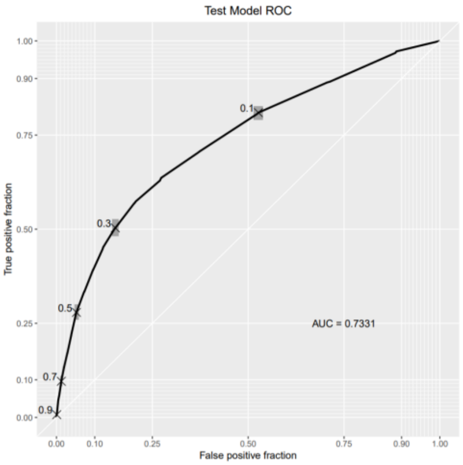

Figure 10. ROC curve for K nearest neighbor.

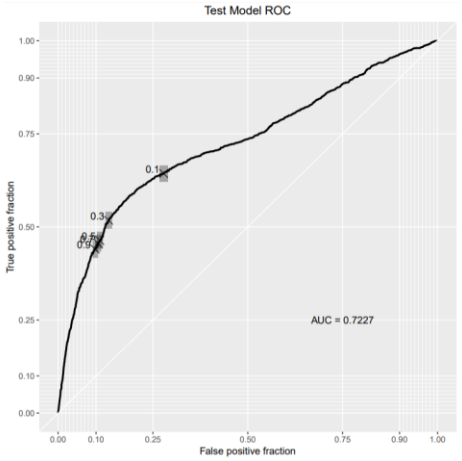

Figure 11. ROC curve for the Naïve Bayes model.

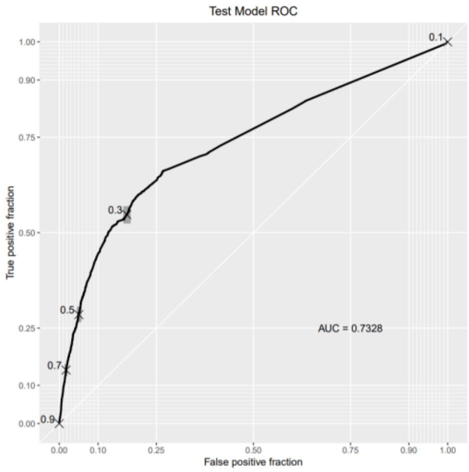

Figure 12. ROC curve for the stacked ensemble model.

Model Evaluation matrix					
Model	Performance metrics				
	Accuracy	AUC	Sensitivity	Specificity	Kappa
Logistic Regression	0.8089	0.761347	0.3008	0.9496	0.308
Decision Tree	0.807	0.733371	0.31724	0.94269	0.3139
K-nearest neighbor	0.8025	0.733062	0.27905	0.94754	0.2805
Artificial neural networks	0.8083	0.762837	0.31141	0.94592	0.3132
Naïve Bayes	**0.7976**	**0.72274**	**0.4637**	**0.8901**	**0.3725**
Ensemble	0.8055	0.73279	0.28541	0.9496	0.291

Table 9. Model Evaluation Matrix.

Figure 7 to Figure 12 show the ROC curves for the models evaluated and table 9 presents the summarized results. Accuracy levels are very similar across models while the probability of correctly predicting that a customer is going to default is significantly higher for the Naïve Bayes model, which also has the highest kappa with a value of 0.3725 that indicates that the model performs 37.25% better than a random guess.

From a business point of view, it is assumed for matter of selecting the best model that the most important thing is to correctly predict when a customer is going to default on his/her credit card in order to minimize the risk. This strategy is also known as a conservative approach. Hence, it is concluded that the best model is the Naïve Bayes.

8. Extra Credits

Given that a high imbalance was found in the data, one way to improve the models computed in section 6 and evaluated in section 7, is to apply up-sampling techniques to the training data set. In this case, random up-sampling was applied to the training data set, aiming to account for the imbalance in the data and to assess if such a strategy would result in an improvement of the evaluation parameters. Table 10 shows the results for the evaluation parameters when random up-sampling was applied to the data training set:

Model Evaluation matrix with random up-sampling					
Model	Performance metrics				
	Accuracy	AUC	Sensitivity	Specificity	Kappa
Logistic Regression	0.7465	0.761495	0.6355	0.7772	0.3561
Decision Tree	0.7552	0.754853	0.6329	0.7891	0.3694
K-nearest neighbor	**0.6725**	**0.724263**	**0.6499**	**0.6788**	**0.2542**
Artificial neural networks	0.7473	0.761378	0.6239	0.7815	0.3527
Naïve Bayes	0.783	0.722806	0.5501	0.8475	0.3836
Ensemble (Gradient Boosting Machine)	0.735	0.750024	0.6265	0.765	0.3342

Table 10. Model Evaluation Matrix with random up-sampling.

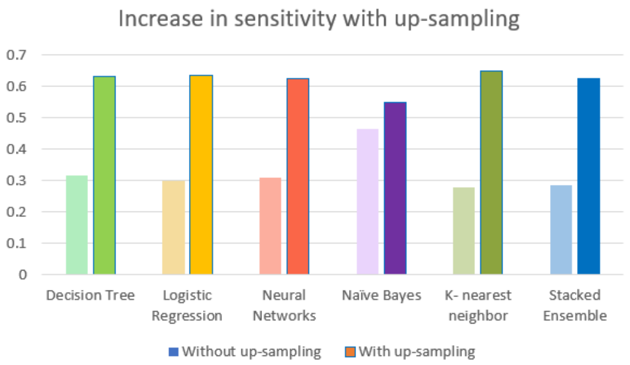

Figure 13. Increase in sensitivity using random up-sampling

Table 10 and Figure 13 show that K nearest neighbor is the model that presents higher probability of predicting correctly customers that default, which the focus in the conservative approach of risk minimization.

9. Real-world insights

The importance of the banking and financial industry for the productivity and profitability of the entire economy is of great importance [10]. In this sense, the health of the industry serves as a guarantor of a healthy economy. However, recent indiscriminate credit card spending has been found to increase the risk of credit default by consumers [6]. Massive high levels of bankruptcy rates of individual credit card holders have the potential to destabilize the credit market [9], a significant component of the financial industry, jeopardizing the stability of the entire economy. A data driven approach to assess risk by predicting credit card default can help to prevent such scenarios.

The main takeaway from the project is that using the model with an up-sampling strategy, the probability of correctly predicting when a credit card user is going to default based on the attributes presented in section 3 about the customer's personal characteristics and the usage behavior in the first six months is 0.6499 percent. On the other hand, the best probability of correctly predicting default without up-sampling was found using the Naïve Bayes model with a value of 0.4637. This demonstrates that accounting for imbalance in the data is crucial to improve the performance of the models.

In addition, accuracy by itself is not a sufficient evaluation metric to evaluate prediction models, especially when the data is highly imbalanced. Also using an up-sampling approach to account for imbalance of the data improves the probability of correctly predicting credit card default by 40.15 percent.

10. Lessons learned

By developing this project, we have learned how the theoretical data analytics and machine learning models can have an application and impact in real life scenarios and how data driven approaches can help in obtaining better process outcomes in different areas, such as financial credit risk management. We have also gained skills in understanding the data analytics methods, their functioning and application. In addition, computational power is an important element to consider when running the models presented in this project.

Another of the learned lessons has to do with the application of decision trees. Although, they are good for visualization, a higher training accuracy does not necessarily translate to a good prediction result.

If having the opportunity, we would have made the effort to use different software languages and not just R. Furthermore, more additional information regarding the cost of wrongly predicting default and the cost of wrongly predicting not defaulting would have been useful to evaluate the models more accordingly to the real-life credit risk management practice.

REFERENCES

[1] Basnet, H.C. and Donou-Adonsou, F. 2016. Internet, consumer spending, and credit card balance: Evidence from US consumers. *Review of Financial Economics*. 30, (2016), 11–22. DOI:https://doi.org/10.1016/j.rfe.2016.01.002.

[2] Cook, J.A. 2017. ROC curves and nonrandom data. *Pattern Recognition Letters*. 85, (2017), 35–41. DOI:https://doi.org/10.1016/j.patrec.2016.11.015.

[3] Dreiseitl, S. and Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*. 32, (2002), 352–359. DOI:https://doi.org/10.1016/S1532-0464(03)00034-0.

[4] Elliehausen, G. and Hannon, S.M. 2018. The Credit Card Act and consumer finance company lending. *Journal of Financial Intermediation*. 34, (2018), 109–119. DOI:https://doi.org/10.1016/j.jfi.2018.01.007.

[5] Gupta, A.. et al. 2017. *Industrial Automation and Robotics*. Mercury learning and information.

[6] Kamil, N.S.S.N. et al. 2014. Examining the Role of Financial Intelligence Quotient (FiQ) in Explaining Credit Card Usage Behavior: A Conceptual Framework. *Procedia - Social and Behavioral Sciences*. 130, (2014), 568–576. DOI:https://doi.org/10.1016/j.sbspro.2014.04.066.

[7] Kim, H. and Devaney, S.A. 2001. *The Determinants Of Outstanding Balances Among Credit Card Revolvers*.

[8] Li, Z. et al. 2017. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today*. (2017). DOI:https://doi.org/10.1016/j.cattod.2016.04.013.

[9] Lopes, P. 2008. Credit Card Debt and Default over the Life Cycle. *Journal of Money, Credit and Banking*. 40, 4 (2008), 761–790.

[10] Patil, S. et al. 2018. Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. *Procedia Computer Science*. 132, (2018), 385–395. DOI:https://doi.org/10.1016/j.procs.2018.05.199.

[11] Williams, G. 2011. *Data Mining with Rattle and R*. Springer.

[12] Yeh, I.-C. and Lien, C.-H. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*. 36, (2009), 2473–2480. DOI:https://doi.org/10.1016/j.eswa.2007.12.020.