A data-driven approach to predict credit card default

Nicolás Navarro Navarro
Hazem Sharaf

VIRGINIA TECH

Problem Statement

- Credit card usage among consumers in the United States has increased over time going from being a product mostly used by high income individuals to a massive adopted product used even by **high risk** individuals

- The purpose of this project is 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 by applying data mining techniques.

- 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.

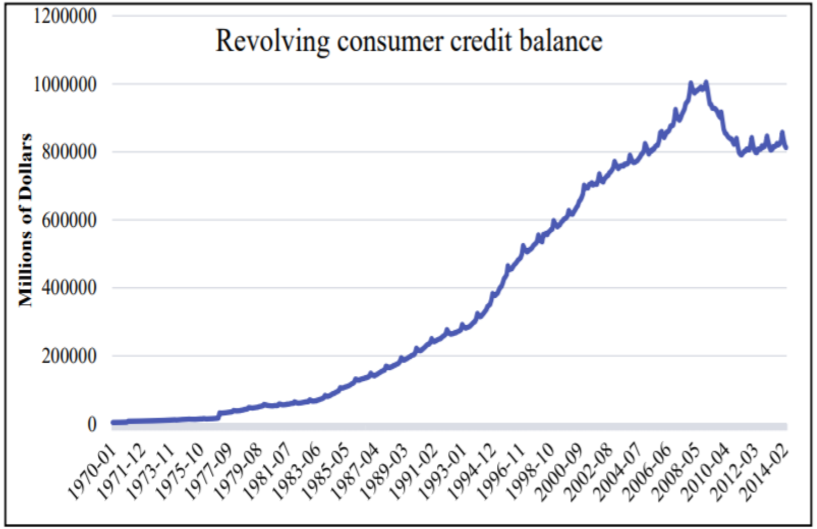


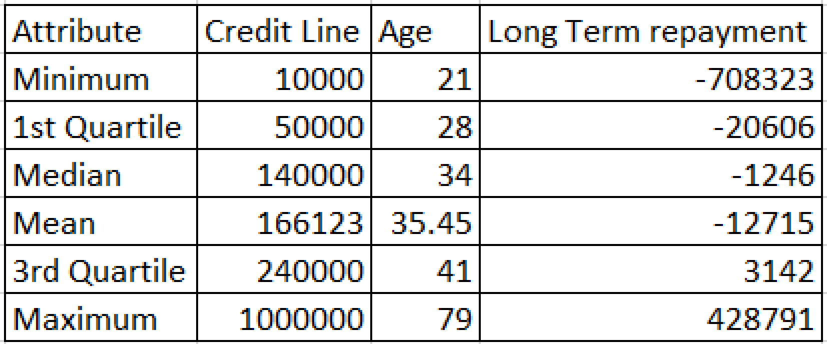
Figure 1. Revolving consumer credit balance from 1970 to 2014.

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:

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.

- **Age: Integer, ranges from 21-79.**
- **Gender: Factor: Male and female**
- **Marital status: 4 levels (including 0)**
- **Education level: Factor: 7 levels**
- **Output Class: Binary: default or not.**
- **Six additional features were added as described later.**



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.

Data Preprocessing

Data adjustment

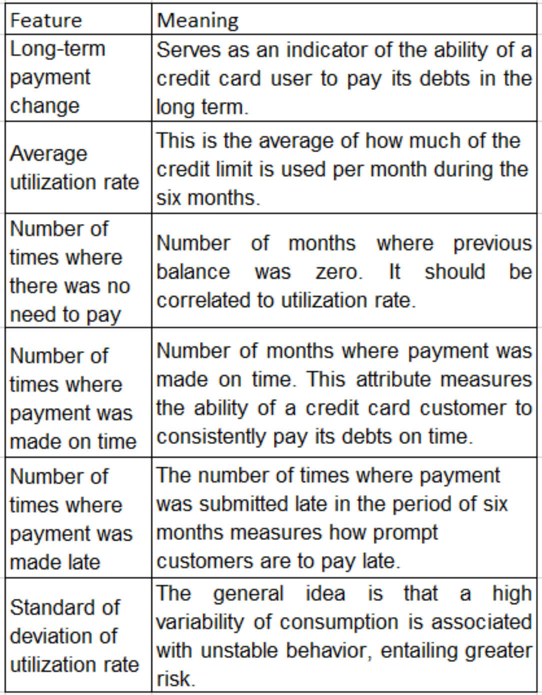
- Education has 4 levels in the original data set, while in the dataset it has 6. Anything above 4 has been set to 4.
- In monthly balance sextuplets, any negative balance (overpayment) is set to zero. (Affects 3932 cells).

Missing data

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

Feature Engineering

- Table 2 presents the new features that were abstracted from the data set.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.



Feature	Meaning
Long-term payment change	Serves as an indicator of the ability of a credit card user to pay its debts in the long term.
Average utilization rate	This is the average of how much of the credit limit is used per month during the six months.
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.
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.
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.
Standard of deviation of utilization rate	The general idea is that a high variability of consumption is associated with unstable behavior, entailing greater risk.

Table 2. Feature engineering

Data Exploration

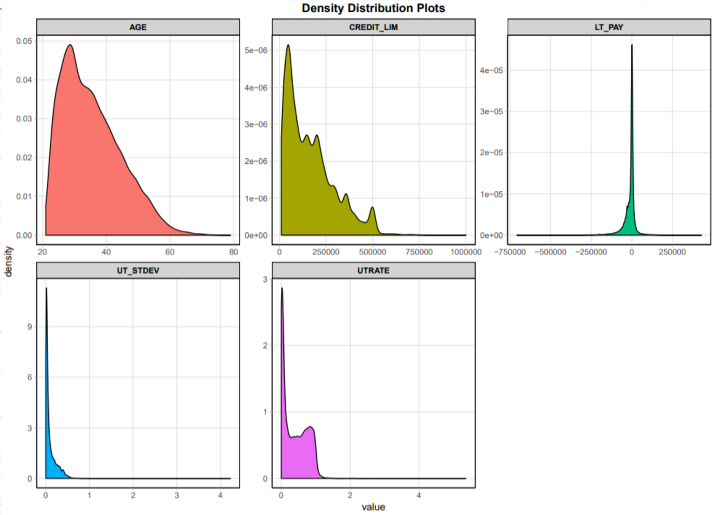


Figure 4. Distribution of the attributes age, credit limit, number of late payments, standard deviation of utilization rate and utilization rate.

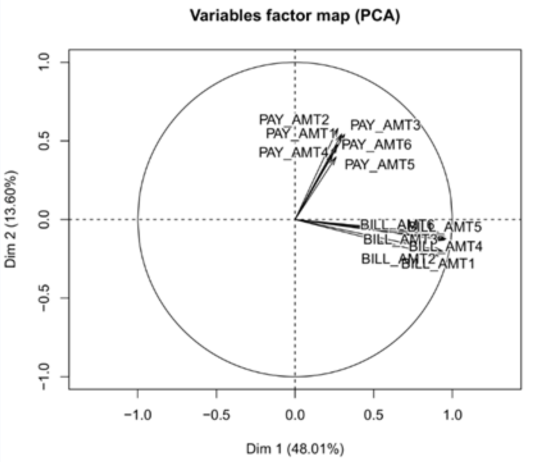


Figure 2. Principal Component Analysis of bill amounts and payments.

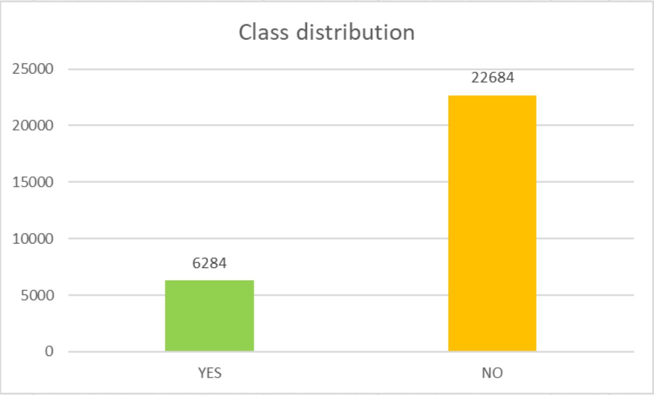


Figure 5. Class distribution in the data set.

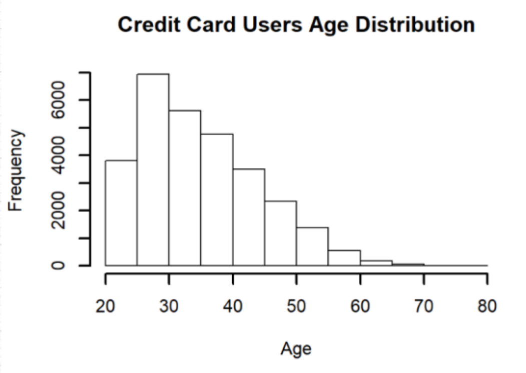


Figure 3. Credit card users age distribution.

Model Building

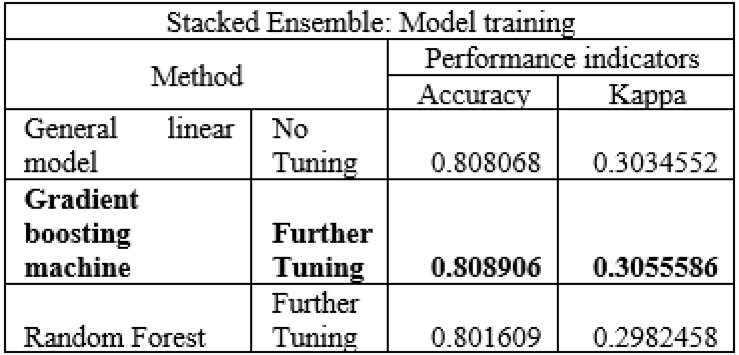
- The approach taken in this project to build the models was to try different values for the parameters that the R packages allowed to modify and see which combination provided the best results. U
- Using the best performing parameters for the training phase, the models were tested using the testing 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.
- 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.

Best Training results for the models		
Model	Performance metrics	
	Accuracy	Kappa
Logistic Regression	0.8085608	0.3093998
Decision Tree	0.8107309	0.3322666
K-nearest neighbor	0.803284	0.2840334
Artitificial neural networks	0.809547	0.320745
Naïve Bayes	0.792782	0.3600722
Ensemble	0.808906	0.3055586

Table 2. General results for the model trained.

Model Building: Stacked ensemble

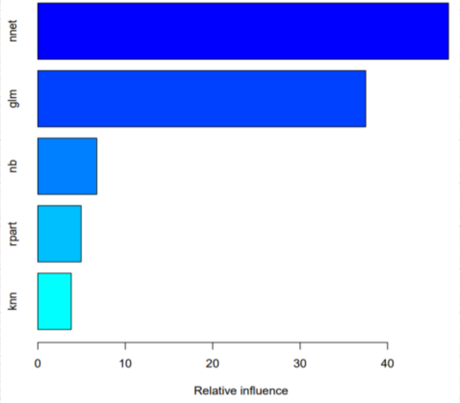
Once all the previous models were trained, a stacked ensemble method was computed having the other methods as the base learners and the ensemble as the meta-learner. Three methods were utilized for this step, a general linear regression model, random forest and gradient boosting machine. The results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

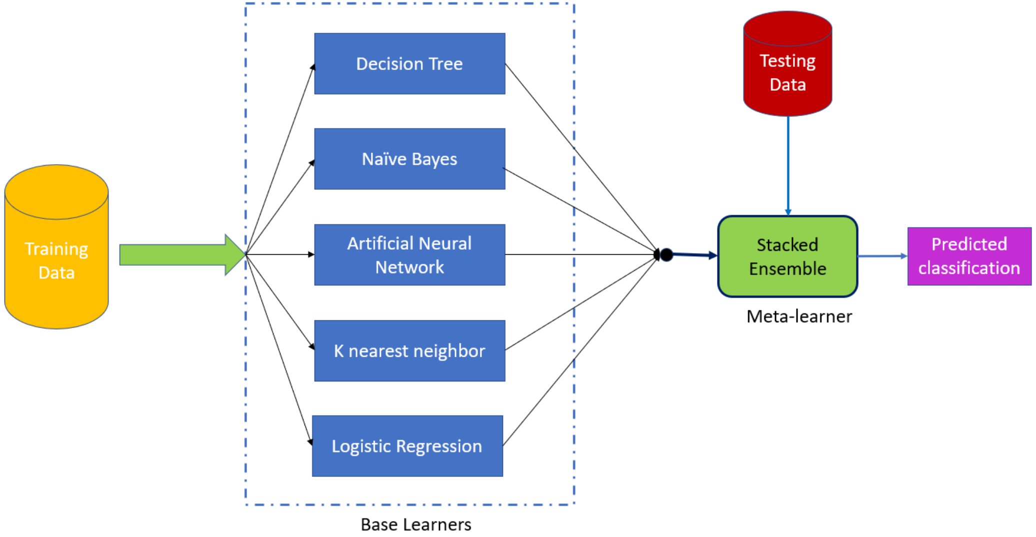


Stacked Ensemble: Model training			
Method		Performance indicators	
		Accuracy	Kappa
General linear model	No Tuning	0.808068	0.3034552
Gradient boosting machine	**Further Tuning**	**0.808906**	**0.3055586**
Random Forest	Further Tuning	0.801609	0.2982458

Table 3. Stacked ensemble training results.

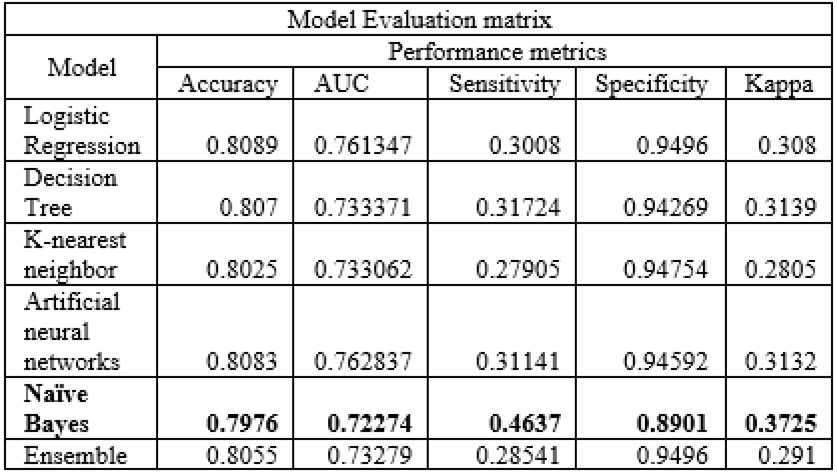
This figure shows the contribution of each of the individual base learners to the stacked ensemble meta-learner. Artificial neural network is the model that contributes the most to the stacked ensemble.





Model Evaluation

- The evaluation of the models is impacted by the imbalance of the data
- In such scenario, the accuracy metric alone does not provide adequate information about the classifier's performance.



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 4. Model Evaluation Matrix

The following is a summary of the methods selected for evaluating the models:

Receiver Operating Characteristic Curve: This curve plots the sensitivity as a function of the specificity, The area under the curve is then taken as a metric for the classifier performance.

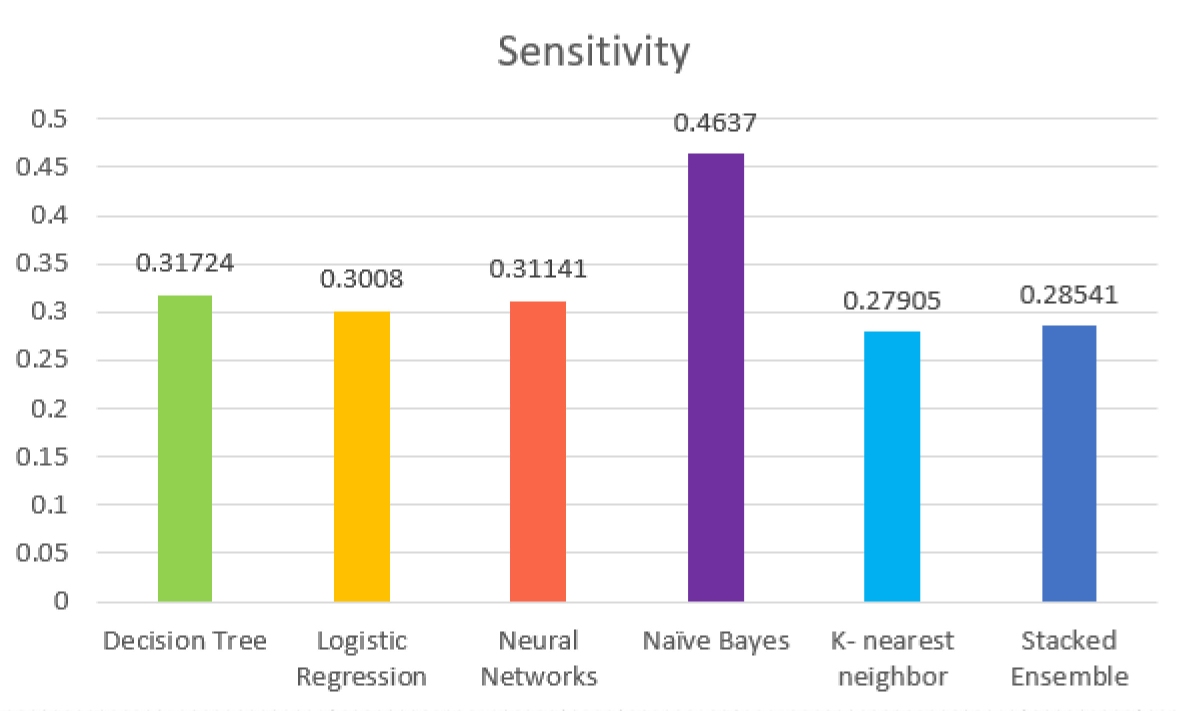
Sensitivity: Probability of correctly predicting that a customer is going to default.

Specificity: Probability of correctly predicting that a customer is not going to default.

Kappa: Reflects the improvement that the model represents over a random guess.

Model Evaluation: Sensitivity visualization

This figure shows the difference in the sensitivity across different models. The probability of correctly predicting default is given a great emphasis in this project, where a conservative approach of minimizing risk is assumed.

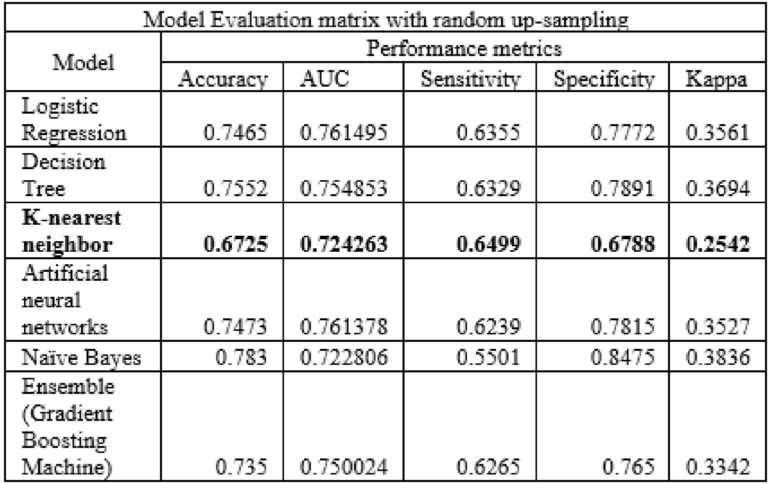


VIRGINIA TECH

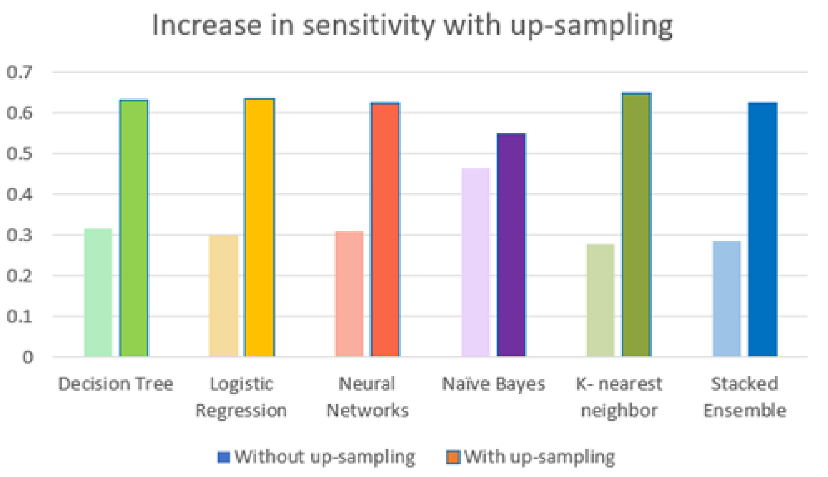
Extra Credits: Accounting for Imbalance

Given that a high imbalance was found in the data, one way to improve the models computed in in section model training and evaluated in model evaluation, 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

The table and figure presented below 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.



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



Real World Insights and main takeaways.

- The importance of the banking and financial industry for the productivity and profitability of the entire economy is of great importance.
- Massive high levels of bankruptcy rates of individual credit card holders have the potential to destabilize the credit market, 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.
- 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.
- The best probability of correctly predicting default without up-sampling was found using the Naïve Bayes model with a value of 0.4637.
- 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.

VIRGINIA TECH