Personal Loan Financing Model in Brazil - Milestone

Group 2
Salman Aamer, Michael Jennings, Ashwin Ringadoo, Alina Shestiaeva

Abstract

The goal of our project is to build a Machine Learning model capable of predicting whether a given loan should be granted to an individual based on an array of features such as work experience, current earnings, etc. We are planning to look at prior transactional data to determine whether a certain individual taking on a loan will be more likely or not to be placed into a state of default or loan delinquency. We are currently targeting the emerging market of Brazil and focusing on individual consumer loans as we believe this is a segment where there is a strong potential for improvement. The data we have sourced includes a lot of information of individual customer's demographics. It is our objective to not blindly run a regression model to only consider to maximize profit or to minimize the error rate, but to build a model that corrects for any biases that may occur naturally through the regression process. As such we will tackle the biased distribution of loans to create a model where gender, ethnicity and age does not become a determining factor of loan granting and adjusts for any model unfairness.

Introduction and Related Work

There has been significant research on potential methods to evaluate the risk of personal loans (credit scoring). One of the most popular methods in credit scoring is to build the prediction model that estimates the probability of the default of the particular client (Peresetsky, Karminsky, Golovan (2011), Zanin, L. (2020)). However, this kind of model has a lot of factors that needed to be taken into account and treated properly. For instance, if the historical information of a certain client is missing or very rare, the logistic regression shows a very poor predictive performance, and in this case, the log-F prior and ridge regression methods are preferred (Ogundimu, E. O. (2019)). Another important topic is that machine learning algorithms can create discrimination based on protective attributes, such as race, color, religion, gender, and disability. Hardt, Price, Srebro (2016) showed that, for example, Bayes optimal non-discriminating (according to our definition) classifier can intuitively solve this issue. Wattenberg, Viégas, Hardt (2016) indicated that correctly chosen "threshold classifier" (when the bank picks a particular cut-off or threshold, and people whose credit scores are below it are denied the loan, and people above it are granted the loan) can efficiently solve the problem. In our research, we are going to build the probability of the default model considering all discriminatory factors into account.

Data and Methodology

We have decided to make use of the publicly available data from the 13th Pacific-Asia Knowledge Discovery and Data Mining conference. The data includes the credit card applications of Brazilian customers in a one year period. The data will include information on acceptance of application, demographics (age, sex, marital status,etc), income, education. The aim of the project is to evaluate the credit scoring model robustness against performance degradation caused by market gradual changes. Data models for credit card scoring at times are based on profit maximization and can lead to an inherent bias or discrimination of individuals from certain demographics. Our aim is to evaluate how a model can be made without a threshold classifier.

0.1 Data Cleaning

Prior to the modelling of credit default rates, a substantial data cleaning process is required. The raw data-set has a number of limitations, which include but are not limited to; i.) variables with a large number of missing observations, ii.) variables that have no variation, iii.) variables that are represented numerically but are intended to capture categories, and iv.) variables with invalid values. Table 1 in the Appendix summarises the data cleaning process for each variable.

We now briefly describe the procedures in Table 1 in more depth for a number of examples. The variable PAYMENT_DAY indicates which days of the month an individual makes a payment and takes the values 1, 5, 10, 15, 20 or 25. Given that a larger value has no natural interpretation, yet the payment day may have important predictive power, we decide to one-hot encode this feature. RESIDENCE_TYPE takes one of five values, which we also one-hot encode. For AGE, we remove three observations that have values of 3,7 and 14 as it may be unreasonable for an individual of this age to have a credit card. For the variables that are removed due to missing observations, most of these have over half of the observations missing. AGE is split into six age brackets: 17-25, 26-35, 36-45, 46-55, 56-75 and 75 and above. The age age brackets are then one-hot encoded. A small number of observations for SEX have entries of 'N', which are removed as we cannot interpret these. We chose to one-hot encode OCCUPATION_TYPE as a factor representing the employment status of a given individual; while we do not have any information as to what the encoding represents, we found it would be a relevant variable to include in our model. Furthermore, we considered another highly dimensional variable: RESIDENCIAL_CITY, which corresponds to the city of residence of the person sending the application. To account for the number of high categories, we instead decided to consider the top 10 cities with the highest percentage of default (when TARGET_LABEL = 1), but only with cities where there are more than 100 entries. Because of the size of certain cities, it made more sense to choose cities where there is likely significantly higher odds of default.

The raw data has 50,000 observations. Cleaning the data reduced the number of observations to 48,383. The figures below illustrate a number of descriptive statistics of our cleaned data.

”””

		Target Label	
		0	1
Gender	Male	13419	5057
	Female	22339	7568

Table 1: Target Labels by Gender (Cleaned Data)

0.2 Modelling

Our methods rely on logistic and linear regression, KNN, as well as classification trees and ensemble methods. We have performed an initial Grid Search to determine the best set of hyper-parameters and added regularization parameters in order to generate a model that is less prone to over-fitting.

After manually choosing the best set of hyper-parameters and the reasonable features to make the predictions, we also decided to examine Automated Machine Learning and to check if we can implement feature engineering more precisely and to choose better model to predict the probability of default. Using Microsoft Azure's Automated Machine Learning, we can create customized models to our data that can accurately forecast the necessary business outcomes.

Before, adjusting for fairness we evaluated the best predictor model without considering the gender of the individuals. Our analysis is presented in the results below.

However, this selected model (logistic regression) has to be adjusted for fairness to ensure that we are not discriminating based on gender. Our model looks at cross-sectional data on different

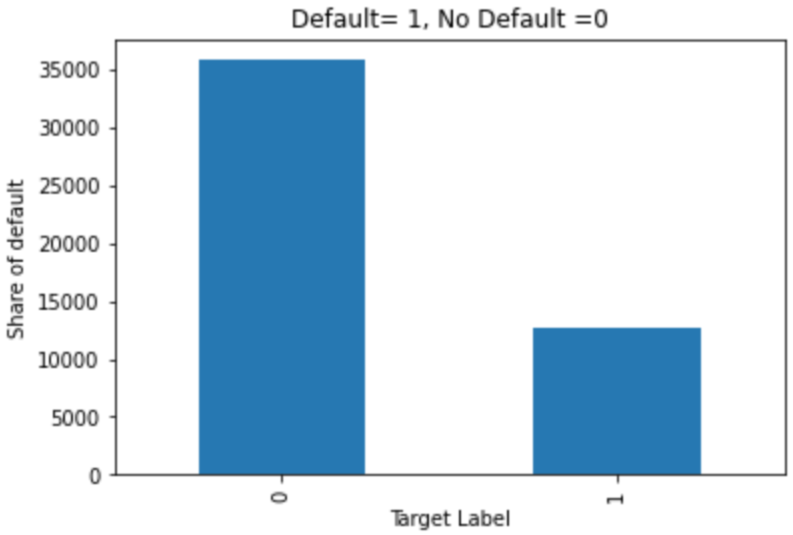

Figure 1: Frequency of Outcomes

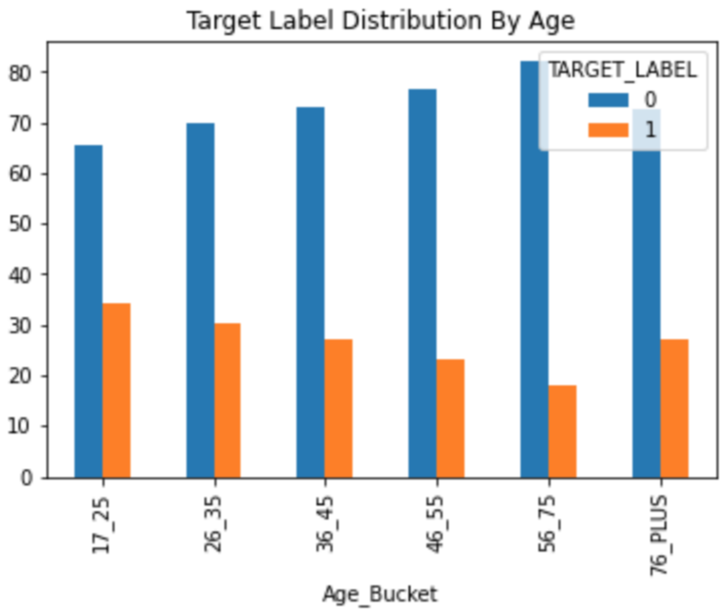

Figure 2: Outcome by Age

demographics to ensure that there is equal opportunity. Specifically, when training our model we specify demographic parity on the protected attribute of sex. Demographic parity requires that individuals are offered the opportunity independent of membership in the protected class. In this context, it ensures that males and females are offered the same types of loans, irrespective of their gender. We apply the same methodology to age by defining an individual over 45 years old as 1 and 0 otherwise. Although the implementation is automated using Microsoft Azure's Fairlearn package, it is useful to better understand what this involves. Some of these details are provided below and readers are referred to fairlearn.org for further technical details. References for the coding implementation are also provided (Github, Unfairness Mitigation).

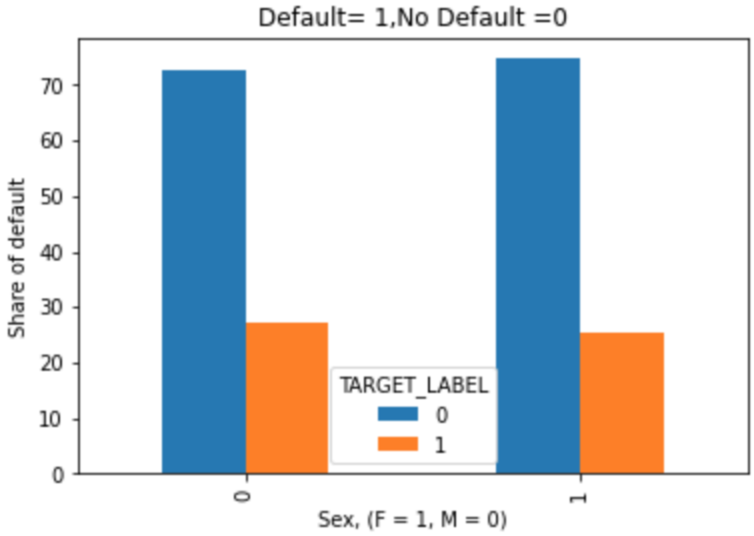

Figure 3: Outcome by Gender

In general, parity constraints require that some aspect of the precictor's behaviour be comparable across groups defined by sensitive features (i.e. gender). For notation purposes, define $\mathbf{X}$ to be the feature vector used for predictions, $\mathbf{A}$ as a single sensitive feature (gender), and $\mathbf{Y}$ as the true label. Parity constraints are phrased in terms of expectations with respect to the distribution over $(\mathbf{X}, \mathbf{A}, \mathbf{Y})$. In the context of a binary classification problem, such as that of logistic regression with binary outcomes, demographic parity is defined as follows: a classifier h satisfies demographic partiy under a distribution over $(\mathbf{X}, \mathbf{A}, \mathbf{Y})$ if its prediction $h(\mathbf{X})$ is statistically independent of the sensitive feature $\mathbf{A}$. Equivalently:

$$E[h(\mathbf{X})|\mathbf{A} = a] = E[h(\mathbf{X})],\ \forall a$$

As such we will know what are the implicit biases included into our model and how we can readjust our training method. It is crucially important as well to determine that a complete accurate model reflecting a 100 percent accuracy is not our objective goal here. The important aspect to consider is that the model should actually predict some defaults (1's in our data), as well as consider the discrimination objective of this research. The goodness of our model will take into account how we can implement demographic parity with respect to features such as sex or age. Splitting the existing data on the training and testing sets, we can evaluate the different metrics that will allow us to check the quality of the model.

Results

0.3 Base models

Our initial analysis was performed on the set of data after removing discriminatory variables including gender, age, marital status. The variables included were standardized income (proxy for being employed), other incomes (proxy for having sufficient funds for repayment in the case of being unemployed), quantity of cars (proxy for personal asset value), months in residence (proxy for being a local resident), possession of credit card (proxy for having a credit history), self-reported employer name (proxy for being employed and ability to perform back ground checks), and type of product applied (encoding is unknown, however we believe it will help our model in choosing the correct parameters).

The analysis performed with the various models led us to select the logistic regression as our preferred model. We have used Grid Search to help choose the optimum parameters.The classification accuracy on the test data for the models used are presented below. As the figures below indicate, the logistic regression leads us to having the greatest precision on both of the target variable values (0 and 1). Our analysis is in line with the literature presented above which highlights the use of a logistic regression in the case of sufficient historical client information (Ogundimu, E. O. (2019)).

We will be analyzing the data using the possible discriminatory variables (age, gender, marital status) to evaluate the true positive rates from our model on these individual variables to evaluate if there is a bias and will then correct for this bias.

The data analysis of our stage involved using multiple machine learning algorithms that would be suitable for a classification problem as presented in this case. The models included a logistic regression, the k-nearest neighbors algorithm (KNN), decision tree algorithm and ensemble learning methods.

As our purpose for the model is to ensure that there is no gender discrimination, we started by first checking the level of default in our data sample without considering gender. Please note that the target variable takes the value of 1, when the individual's debt is considered as bad debt, which would be in the case of a significant delay in making repayments (60 days or more), while it takes the value of 0 in the case when the individual had repaid their debt. The following table 2 represents all the results. For the purpose of the further comparison, according to all metrics the logistic regression was chosen as the best model.

	Precision	**F-1 score**	**Accuracy**
Logistic Regression	0.74	0.85	0.7423
K- Nearest Neighbours (KNN)	0.74	0.85	0.7421
Classification Tree	0.74	0.85	0.7423

Table 2: Precision, Recall, F-1 score and Accuracy for classification algorithms, test set

Using the previous algorithms, two ensemble models were built:a) A model "Mode" that uses a majority voting scheme of the predictions of each algorithm to classify new records. b) A model "Mean" that uses the mean of the probabilities of all the algorithms that the record belongs to the class of interest (i.e., y=1). If the mean is higher than the cutoff value (i.e., p=0.4), the record is classified in the class of interest. If not, it is classified in the other class. Results are presented in the table 3. The cutoff was chosen arbitrarily. The most typical cutoff is 0.5, however since our data does not have a high level of historical defaults, we slightly decreased it by 0.4.

	Precision	**F-1 score**	**Accuracy**
Ensemble Model Mode	0.74	0.85	0.7423
Ensemble Model Mean	0.74	0.85	0.7417

Table 3: Accuracy for the Ensemble models, test set

Given the fact that all chosen models gave almost the same accuracy of approximately 74%, we will consider the Mean Ensemble Model as the best one, since according to the confusion matrix it predicts more true total defaults.

For the purpose of the coherence of the research, we also implemented Automated ML model to possibly transform our features and choose better model according to various metrics.

While working with the selection of the variable, the Auto ML passes all three necessary tests (class balancing detection, missing feature value imputation and high cardinality feature detection) that proves that initially selected variables in the model above make sense. The list of best 10 model according to AUC is presented below 4.

As a results, the Voting Ensemble was chosen as the best one with the accuracy 0.74 (see table 5))

№	Top-10 Models	AUC weighted
1	**SparseNormalizer LightGBM**	0.5799
2	**StandardScalerWrapper DecisionTree**	0.5799
3	**MaxAbsScaler LightGBM**	0.5799
4	**SparseNormalizer LightGBM**	0.5799
5	**RobustScaler LightGBM**	0.5799
6	**MinMaxScaler ExtremeRandomTrees**	0.5799
7	**StandardScalerWrapper LightGBM**	0.5799
8	**MinMaxScaler GradientBoosting**	0.5799
9	**VotingEnsemble**	0.5867
10	**StackEnsemble**	0.5867

Table 4: TOP-10 best model according to Auto-ML model

	Precision	F-1 score	Accuracy
The Voting Ensemble model	0.74	0.75	0.74

Table 5: Accuracy for the Auto ML, test set

0.3.1 Fairness issues

For the purpose of the further analysis of discrimination issues, we will consider the logistic regression as the base model, since the Fairlearn package in Microsoft Azure does not handle the ensemble model and the differences in the predictions of other modeles are not that significant,

We can also check how the chosen model works separately for gender and age groups. The tables below (6 and 7) present the results of our initial model - logistic regression for examined groups.

	Precision	F-1 score	Accuracy
Logistic Regression for men	0.72	0.84	0.72
Logistic Regression for women	0.75	0.86	0.75

Table 6: Precision, Recall, F-1 score and Accuracy for mean and women separately for logistic regression, test set

	Precision	F-1 score	Accuracy
Logistic Regression for people younger (=) than 45	0.71	0.83	0.71
Logistic Regression for people older than 45	0.79	0.88	0.79

Table 7: Precision, Recall, F-1 score and Accuracy for people younger and older than 45 separately for logistic regression, test set

The results from our model adjusted for fairness for gender are presented in the table 8 below. The accuracy for females is 75.3 percent, while the accuracy for males is 72.4 percent. The selection rate which represents the fraction of points from classified as default or bad debt is 0.0224 % for females while it is 0.0179 % for males. This accuracy for females is higher than the accuracy we received from our original model that did not consider gender and led to an accuracy of 74.2 percent.

The results from our model adjusted for fairness for age are presented in the table 9 below. The accuracy for people older than 45 is 79.2 percent, while the accuracy for younger people is 70.4 percent. The selection rate which represents the fraction of points from classified as default or bad debt is 0.0231 % for younger people while it is 0.017 % for older people. This accuracy for older

	Accuracy	Selection rate, %	False Positive Rate, %
Fairness model results for men	0.724	0.0179	0.0247
Fairness model results for women	0.753	0.0224	0.0298
Overall	0.753	0.0207	0.0278

Table 8: Fairness model results for males and females

people is higher than the accuracy we received from our original model that did not consider age and led to an accuracy of 74.2 percent.

	Accuracy	Selection rate, %	False Positive Rate, %
Fairness model for people younger (=) than 45	0.708	0.0231	0.0327
Fairness model for people older than 45	0.792	0.017	0.0215
Overall	0.742	0.0207	0.0278

Table 9: Fairness model results for the age

Conclusion

As a result, we built the Machine Learning model that is capable to predict whether a given loan should be granted to an individual based on the selected features - type of the residence, payment day, type of the product, occupation type, marital status and other variables. We were focusing on the individual consumer loans in Brazil, since we believe that in this emerging market there is a strong potential for improvement of the scoring system. We implemented logistic regression, KNN, Classification Tree and Ensemble Model achieved accuracy of 74% for every model. We chose the Ensemble Model as the most relevant model for our research, since it allows to predict more defaults. The Voting Ensemble Model with the accuracy 74% as well was chosen after implementing Automated Machine Learning. By considering Fairlearn model using Microsoft Azure and adjusting for any model unfairness, we showed that gender and age do not become determining factors of a loan granting.

References

1) Peresetsky, Anatoly A., Alexandr A. Karminsky, and Sergei V. Golovan (2011). "Probability of default models of Russian banks." Economic Change and Restructuring 44.4: 297-334.

2) Ogundimu, E. O. (2019). Prediction of default probability by using statistical models for rare events. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), 1143-1162.

3) Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. № 1610-02413.

4) Wattenberg, M., Viégas, F., Hardt, M. (2016). Attacking discrimination with smarter machine learning. Google Research, 17.

5) Zanin, L. (2020). Combining multiple probability predictions in the presence of class imbalance to discriminate between potential bad and good borrowers in the peer-to-peer lending market. Journal of Behavioral and Experimental Finance, 25, 100272.

6) Github. Unfairness Mitigation with Fairlearn and Azure Machine Learning. Available at: https://github.com/Azure/MachineLearningNotebooks/blob/master/contrib/fairness/fairlearn-azureml-mitigation.ipynb

7) Fairlearn. User guide available at: https://fairlearn.org/main/user_guide/fairness_in_machine_learning

9)The following Microsoft Resource was used in implementing our Auto ML Model https://docs.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train

Github Repository: https://github.com/salmanaamer/6.862.Spring.2021.Group.2.git

Table 10: Summary of Data Cleaning

Data Limitation	**Variables**	**Solution**
Categorical variables with more than two categories	RESIDENCE_TYPE PAYMENT_DAY PRODUCT OCCUPATION_TYPE	One-hot encoding
Variables with invalid values	SEX	Remove observations
Unreasonable values	AGE MARITAL_STATUS	Remove observations
Constant values	EDUCATION_LEVEL CLERK_TYPE QUANT_ADDITIONAL_CARDS FLAG_MOBILE_PHONE	Drop Variable
Too many missing observations	MATE_EDUCATION_LEVEL APPLICATION_- SUBMISSION_TYPE RESIDENCIAL_PHONE- _AREA_CODE PERSONAL_ASSETS_VALUE PROFESSIONAL_STATE PROFESSIONAL CITY PROFESSIONAL_BOROUGH PROFESSIONAL_PHONE- _AREA_CODE MATE_PROFESSION_CODE FLAG_HOME_ADDRESS- _DOCUMENT FLAG_RG FLAG_CPF FLAG_INCOME_PROOF FLAG_ACSP_RECORD	Drop variable
No clear interpretation of encoding	NACIONALITY	Drop variable
Duplicate variable	QUANT_SPECIAL- _BANKING_ACCOUNTS	Drop variable
Continuous variables used for discrimination testing	AGE	Divide into 6 age buckets and one-hot encoded
Other Geographic variables	STATE_OF_BIRTH CITY_OF_BIRTH RESIDENCIAL_STATE RESIDENCIAL_CITY RESIDENCIAL_BOROUGH PROFESSION_CODE PROFESSIONAL_ZIP_3 RESIDENCIAL_ZIP_3	Drop variable. The RESIDENCIAL_CITY variable has been chosen to represent geographic location and adding other geographic variables will likely add colinearity to the model.
Similar variables	PERSONAL_MONTHLY- _INCOME OTHER_INCOME	Summed and then standardised
Geographic Variable	RESIDENCIAL_CITY	Created 10 binary variables for the top 10 cities with the highest default rate (for cities with > 100 observations)