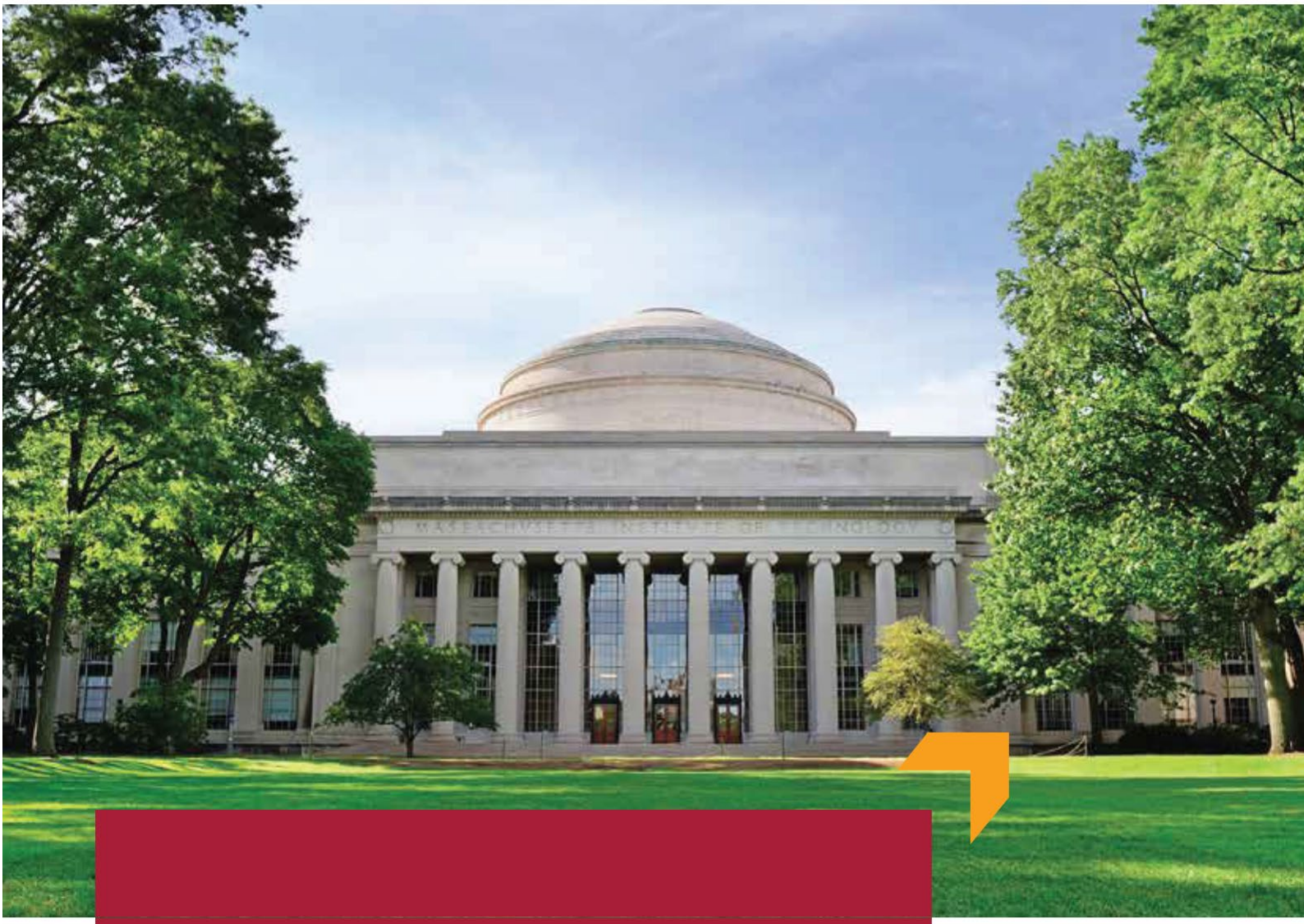
PROFESSIONAL EDUCATION



APPLIED DATA SCIENCE PROGRAM

Capstone Project: Loan Default Prediction

Author: Phillipe Vilaça Gomes

CONTENT

1. Executive Summary

1.1 Context

- Home loans are one of the most crucial credit lines to increase banks' profits. As defaulters can reduce drastically the mentioned profit, banks often establish a process to avoid bad loans (NPA).
- In this process, the bank checks a considerable variety of attributes and came up with a decision. In this way, many banks are trying to automate this process in a free of bias and efficient way.
- Therefore, this project proposes a model based on Machine Learning techniques and it is intended to automate the decision-making process of approving loans by recognizing customers likely to default.
- The project was carried out in three stages:
    - Step 1: Exploratory Data Analysis.
    - Step 2: Model Comparison.
    - Step 3: Decision-Making in the Model.

1.2 Key Takeaways

- From the **Exploratory Data Analysis**, some of the features have presented a remarkable pattern for loan default.
    - About 9% of the clients with *debt-to-income ratio* information in the loan application have defaulted on the loan.
    - 93% of customers with a *debt-to-income ratio* lower than 25 can afford the loan.
    - Customers with a high *debt-to-income ratio* tend to become defaulted on loans.
        - Roughly 50% of customers with a *debt-to-income ratio* higher than 43 have defaulted on the loan.

        - Customers with a d*ebt-to-income ratio* higher than 45 have always defaulted on the loan.
    - About 62% of the clients missing d*ebt-to-income ratio* information have defaulted on the loan.
    - Applications from the *sales industry* have a stronger impact on delinquency.
    - Customers whose *number of derogatory reports* is more than 3 have more chances of defaulting.
    - The information about the *current value of the property* also presents strong importance to loan default.
        - About 94% of the clients missing this information have defaulted on the loan (against 19% of the clients with this information).
    - The *reason why* the loan is required does not interfere with default.
- From the **Proposed Model**, **Random Forest with Tuned Hyperparameters**, the most important features to consider in the decision-making on loan applications are:
    - Debt-to-income ratio.
    - Miss the Debt-to income ratio information in the loan application.
    - The age of the oldest credit line.
    - The amount of loan requested.
    - The current value of the property.
    - The amount due on the existing mortgage.
- With the proposed model:
    - 75% of customers who defaulted on loans were correctly identified as not eligible customers.
    - 91% of clients who duly paid the loans were correctly identified as eligible customers.
    - The model was able to correctly classify 88% of customers.

1.3 Final Proposed Model Specification

- The proposed model was the **Random Forest with Tuned Parameters**. This algorithm is a bagging approach where the base models are decision trees. In this way, the algorithm uses a subset of the features, chosen randomly, for each node's branching possibilities. Thus, Bootstrapped samples are taken from the original training data, and on each bootstrapped training dataset, a decision tree is built by considering only a subset of features at each split. The results from all the decision trees are combined and the final prediction is made using, in this case, voting.
- The *GridSearch* algorithm was used to perform hyperparameter tuning. This is a heuristic algorithm that minimizes the loss of the model based on the values of hyperparameters. The main motivation for using this algorithm is that the mentioned minimization problem presents a non-convex and non-linear nature, which is hard to solve by conventional algorithms that ensure optimality. The following hyperparameters were used in the *GridSearch* algorithm:
    - The number of trees in the forest (100, 250, and 500).
    - The minimum number of samples required to be at a leaf node (1, 2, and 3).
- The best solution obtained by the GridSearch Algorithm was:
    - Number of trees in the forest: 250
    - The minimum number of samples required to be at a leaf node: 3.
- 70% of the dataset was separated for training and the remaining 30% was separated for testing the model.

1.4 Potential Benefits from the Proposed Solution

- The models tested in this project can make two types of wrong predictions:
    - Predicting an applicant will default on a loan when the applicant doesn't default (False positive).
    - Predicting an applicant will not default on a loan when the applicant actually defaults (False negative).
- In the first case, the bank will lose the opportunity to increase its profits from the loan. In the second case, the bank will lose out on the loan. Therefore, the most

important thing to avoid is false negatives, that is, predicting an applicant will not default on a loan when the applicant actually defaults.

- In this way, the bank would want the Recall to be maximized, the greater the Recall, the higher the chances of minimizing false negatives. Hence, the focus should be on increasing the Recall (minimizing the false negatives) or, in other words, identifying the true positives (i.e. not eligible customers) very well, so that the company can provide incentives to control the default rate especially.
- The final solution approach is the one with the best performance over the test Set (30% of the population) based on the maximization of the Recall. In this way, the automatization of the process may allow a significant reduction in the loan default. Besides, it will also allow a better understanding of the important features to consider while approving a loan.

1.5 Key Next Steps

- The proposed model had a remarkable performance and low error with observed data. However, it is subject to a few limitations, including a lack of consideration for credit score (explicitly, besides the number of delinquent credit lines and derogatory reports) and employment history. In this way, it is recommended that stakeholders consider these new features when improving the loan default prediction model.

2. Problem and Solution Summary

2.2 Solution Design

- In this project, 8 different models were tested to solve the Loan Default Prediction Problem. As the target, which indicates if a client has defaulted on a loan or not, is a binary variable, Logistic Regression was selected to be explored in three different versions, as follows:
    - **Model 1**: Logistic Regression.
    - **Model 2**: Logistic Regression with Balanced Precision-Recall.
    - **Model 3**: Logistic Regression with Scaled Feature.

  - **Model 4**: Logistic Regression with Scaled Feature, Balanced Precision-Recall.
- Besides, 2 different models were built from predictive modeling techniques based on Decision Trees:
  - **Model 5**: Decision Tree with Default Hyperparameters.
  - **Model 6**: Decision Tree with Tuned Hyperparameters.
- Finally, to establish an important tradeoff between performance and interpretability, 2 different models based on Random Forest were also selected to be explored as follows:
  - **Model 7**: Random Forest with Default Hyperparameters.
  - **Model 8**: Random Forest with Tuned Hyperparameters.
- The dataset has 5960 loan applications with 13 inputs. One of these inputs is related to the classification in loan default or loan repaid and it is the **target**, all the remaining inputs are the features used to take the classification in eligible (loan repaid) and not eligible (loan default).
- To establish the solution design, the original dataset is split into two parts: The training Set and Test Set.
  - Training set: 70% of the data.
  - Test set: 30% of the data.
- All models were explored using only the Training Set while the Test Set will be used to access their performances. In this way, their performances were assessed using 4 different attributes briefly explained as follows:
  - **Accuracy**: The fraction of classifications (eligible and not eligible clients) that the model has got correctly.
  - **Precision**: The average between the fraction of clients correctly classified as not eligible and clients correctly classified as eligible clients.
  - **Recall:** The average between the fraction of not eligible clients classified as not eligible and the fraction of eligible clients classified as eligible.
  - **F1-Score**: Weighted average of precision and recall.

- A valid solution should have more than 80% in all attributes. As mentioned before, Recall is the most important attribute, and the final decision must consider special attention to this.

2.3 Analysis and Key Insights

- Figure 1 presents the performance of the models when applied to the unseen test dataset data.

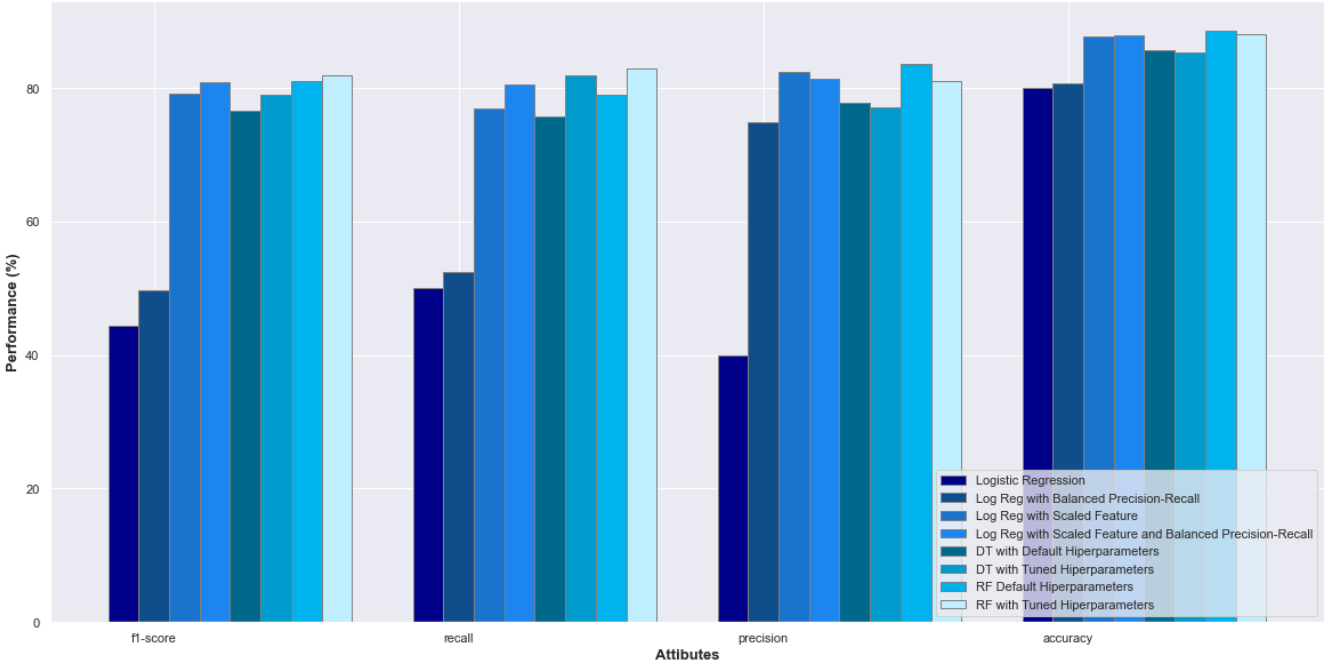


Figure 1: Performance of the models against the unseen test dataset.

- These values are also presented in Table 1 for convenience. Highlights are done for Model 8, which presents the best performance in almost all attributes, and for Recall, the most important attribute in this project.
- Model 8, Random Forest with Tuned Parameters, presents the higher Recall, the higher F1-score, the second higher accuracy, and an acceptable precision of about 81%. For that reason, this model was chosen to tackle the Loan Default Prediction problem.
- According to the proposed Model, the five most important features when conducting the classification procedure are presented in Figure 2.

Table 1: Results for the different attributes.

Model	1	2	3	4	5	6	7	8
Precision	40.01	74.90	82.35	81.37	77.78	77.17	83.58	81.02
Recall	50.00	52.48	76.89	80.50	75.67	81.84	79.03	82.99
Accuracy	80.03	80.64	87.75	87.97	85.62	85.40	88.64	88.08
F1-score	44.45	49.75	79.13	80.92	76.63	79.04	80.98	81.94

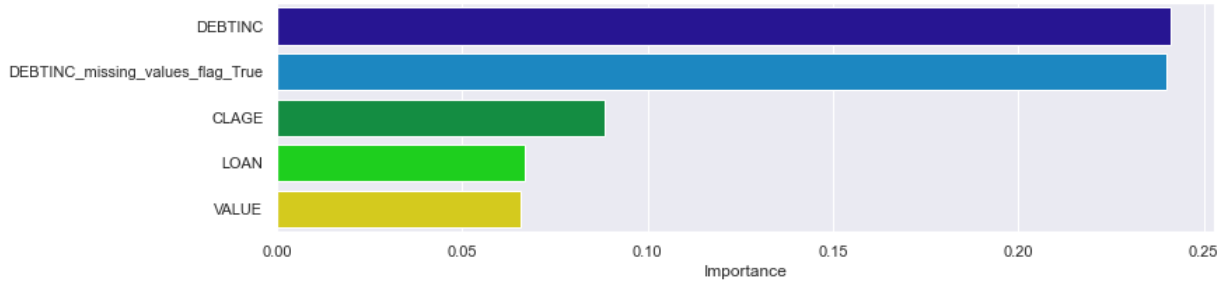


Figure 2: The five most important features to perform the classification in the proposed model.

- DEBTINC is the debt-to-income ratio, CLAGE is the age of the oldest credit lines, LOAN is the amount of loan approved and VALUE is the current value of the property.
- With the proposed model, 91% of the customers that do not have defaulted on the loan were classified as eligible, and 71% of the customers that have defaulted on the loan were classified as not eligible.
- Even though the proposed model – Random Forest with Tuned Parameters - is less interpretative than the Decision Tree with Tuned Hyperparameters, which also had a good performance in all attributes, the remarkable performance of the proposed model compensates for the reduced level of interpretability.

3. Recommendations for Implementation

3.1 Key Recommendations

- The bank should **not approve** loans to those customers with a debt-to-income ratio higher than 40, as those customers have very high chances of defaulting. For those

customers, the bank should consider **approving** the loan only if the age of the oldest credit line is superior to 285 months.

- The bank should **pre-approve** loans to those customers with a debt-to-income ratio of less than 25, (use the proposed model for the final decision) as those customers have very high chances of repaying.
- The bank should **avoid approving** loans in applications missing the debt-to-income ratio.

3.2 Key Actionables for Stakeholders

- Missing value was identified as a characteristic of great importance in the decision-making of loan approval. In particular, missing values for the debt-to-income ratio and the current value of the property appeared as the main factors in several models tested.
- Thus, stakeholders should promote the importance of this information during the application for the loan to better identify customers with higher chances to default while improving the forecast model.

3.3 Expected Benefits

- The proposed model was able to classify 91% of consumers who duly paid the loan as eligible consumers for the loan, and on the other hand, it was also able to classify 71% of consumers who defaulted on the loan as not eligible consumers.
- Thus, it can be concluded that the marginal loss of loan opportunities to eligible consumers is offset by the savings realized from not granting loans to the not eligible consumers.

3.4 Key Risks and Challenges

- Although the proposed model has effectively identified eligible and ineligible consumers for the loan, the model does not guarantee 100% accuracy. Thus, the model should be used to assist in decision-making.
- Finally, the proposed model will be more effective when used in loan applications with the same features considered in this study.