PERSONAL PROJECT

PREDICT THE STATUS OF LOAN

Developer: Marcos Matheus de Paiva Silva

Barra Mansa - RJ

March de 2022

Contents

1 OBJECTIVE 2

2 PROPOSED SOLUTION 3

3 RESULTS 4

4 CONCLUSION 7

BIBLIOGRAPHY 8

1 Objective

In the current project, we fully contemplate the stages of a data science project, so that from the data collected we can probe some relevant questions that would help banks and insurance companies understand about:

- The pattern of People who generally have good conditions to positively receive a loan;
- Predict which customers may or may not receive loans, based on information about customer conditions;
- Which variables in our dataset most influence loan eligibility.

Finally, we provide an online application that allows anyone to predict whether or not an individual is eligible to receive a loan (Deploy). Such a procedure may be performed, based on information either collected or reported by the applicant such as gender, marital status, number of dependents, education, type of employment, applicant's income, co-applicant's income, Amount of loan requested, number of installments , credit history and place of residence..

2 Proposed Solution

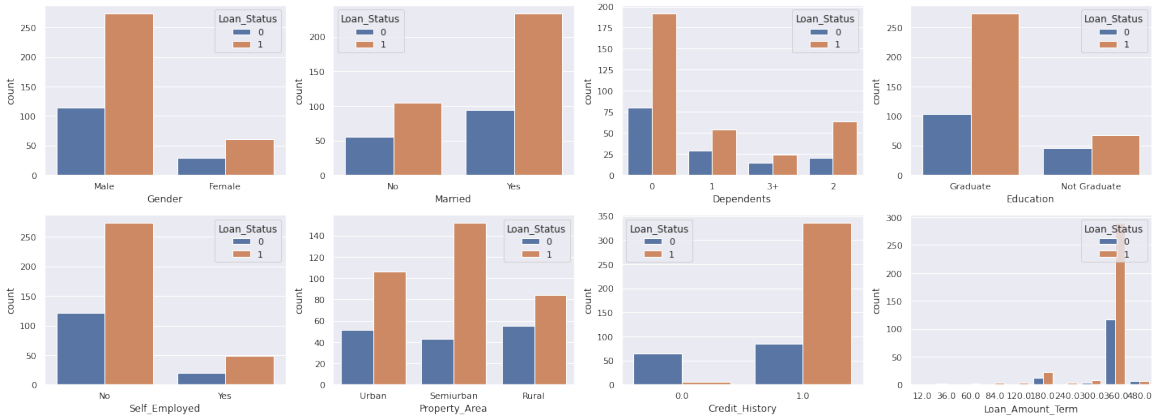
This project was developed following the steps in (GéRON, 2019), and also based on some codes found in (MüLLER; GUID, 2016). To find the answers to the questions raised in the previous section, we first used the Google Colaboratory so that any individual can run the code in python and also better understand all the theoretical argument, insights and methodology used in the project. In Google Collaboratory, libraries to operate with machine learning, data analysis, graphing and files were pandas, sklearn, matplotlib, numpy, seaborn, imblearn.

Furthermore, after creating the machine learning model in Google Collaboratory, the library pickle was used to serialize the model and write it in a file. With the "trained_model.sav" file in hand, we implemented and locally tested the web application using the python language distribution anaconda and with the help of API streamlit we produce the app in a few lines of code. Finally, we host the web app on the Herok App platform where you can test our loan eligibility prediction app.

3 Results

Our Google Collaboratory project went through several steps both to obtain initial insights and to understand and explore our dataset provided by Kaggle. In the data visualization stage, we plot a graph:

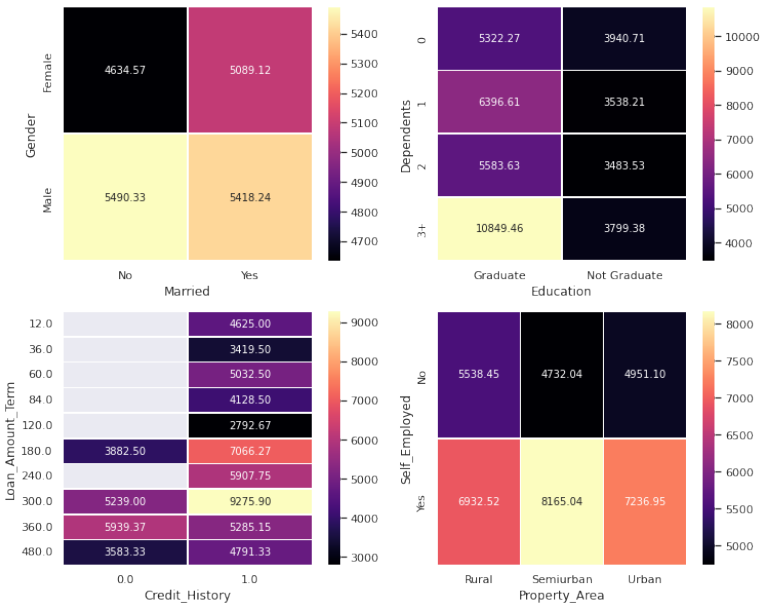
Figure 1 – Frequency x Category Graphic



Source: Author

which seems to indicate that married, childless, college-educated, non-self-employed men who live in semi-urban areas, who have a good credit history, and who make payments within 360 days are reportedly more likely to receive loans. We also draw a heat map that relates categorical data and the applicant income:

Figure 2 – Categories and Applicant Amount Heat Map

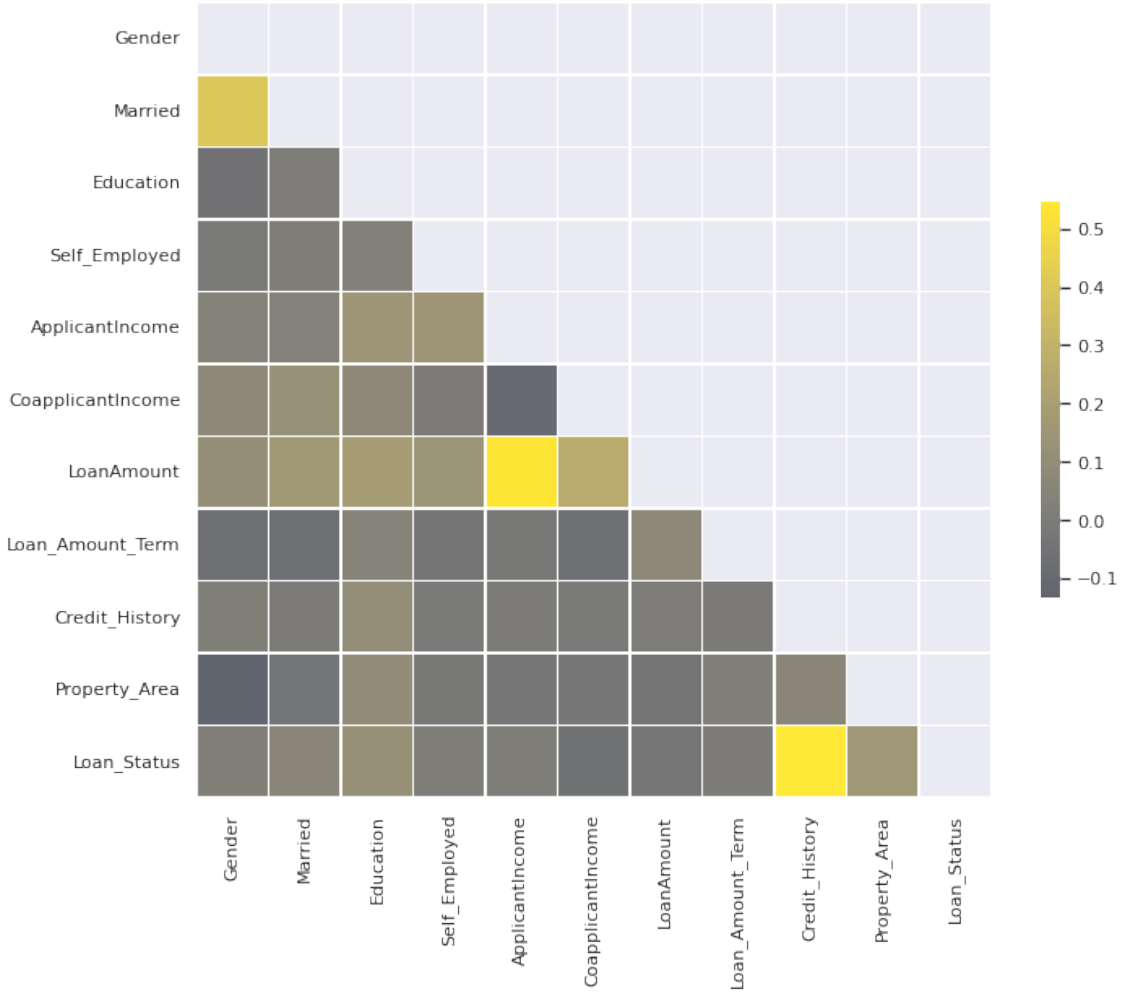


Source: Author

interestingly, this graph reaffirmed some previously stated precepts, in addition to telling us that people with income below the average $\sim$ ₹ 5.255,00 are more likely to receive a loan.

In the exploratory data analysis and feature engineering part, we try to create new data and verify correlations between variables. Therefore, using a correlation tool, we understand that the loan situation has a strong correlation with the credit history and that the amount of loan requested also has a strong correlation with the applicant's income, that is, the higher the individual's salary, the greater the how much credit he requests. Such information can be easily verified in the following chart:

Figure 3 – Correlation map between categorical variables

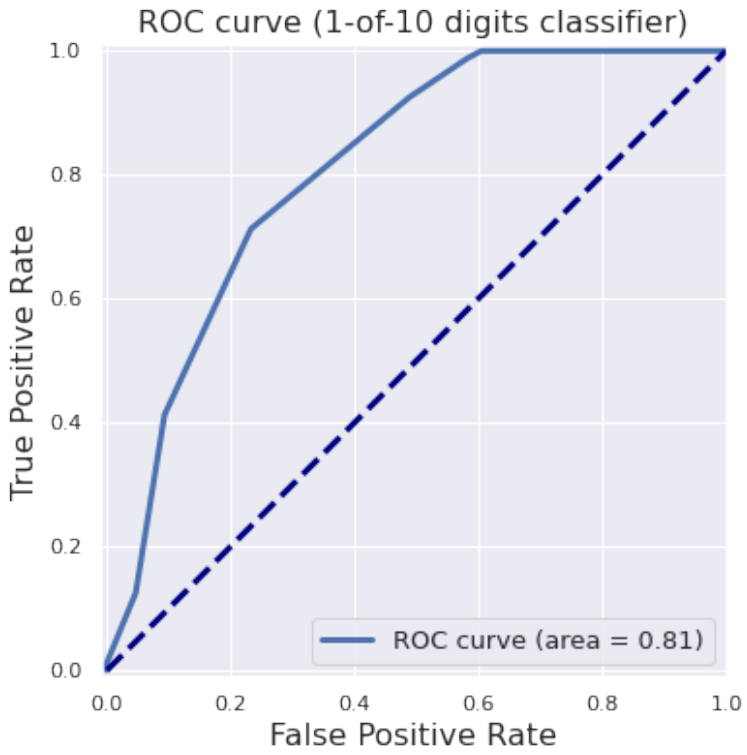


Source: Author

, because we understand that the closer to yellow the greater the correlation between the variables and the closer to gray the lower the correlation. Furthermore, we tried to create new variables, however, the creation of some variables did not have much influence on the final model, and other variables harmed the model, as it included some divisions by zero in it. Therefore, due to these reservations, it was decided not to use new data.

Therefore, in order to predict whether an applicant is eligible to receive a loan, it was decided to implement the model by applying the **k-nearest neighbors** algorithm, where applicants are classified according to the plural vote of their neighbors, in which each object is assigned the most similar class among its k nearest neighbors. For example, if individuals have a lot of similar information such as applicant's income and co-applicant's income they will be classified in the same group, otherwise, individuals are classified as being in another group. This is done by voting by the 15 neighbors with those most similar values of applicant's income and co-applicant income, that is, if we have 15 close neighbors with similar values, they will be classified as being in the same group.

Finally, in our evaluation we used the ROC curve and the AUC value. The ROC curve describes the rate of true positives by the rate of false positives. In our case, we aim to have a high rate of true positives, as we would like to know who the people who should receive the loan are. The AUC value ranges from 0 to 1, and the closer to 1, the better our estimates.



Source: Author

As in the figure we had an area of 0.81, ie the AUC value close to 0.81 we had a good performance.

4 Conclusion

In the course of this project, we developed relevant concepts in data science and analysis, in order to suggest the solution to the problem of credit eligibility of an applicant that can be freely used by any bank or insurance company. We go through several steps precisely in order to extract as much information as possible from our dataset. As a result, we found that married, childless, university-educated, non-self-employed men who live in semi-urban areas, who have a good credit history and who make payments within 360 days are supposedly more likely to receive loans.

Furthermore, we employ several techniques that can be verified in google collaborative, in order to further improve our model and increase its efficiency, so we found an area under the ROC Curve of 0.81. Finally, we implement our online model and make it available for anyone to use as we need to verify an applicant's eligibility in real time. For all these caveats, we understand that a way to improve our model in the future would be to use a much larger dataset, and allow the model to be saved and update itself to learn more and more, as new data is entered into the algorithm.

Bibliography

GéRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2ª. ed. GCanada: O'Reilly, 2019. Citado na página 3.

MüLLER, A. C.; GUID, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1ª. ed. United States of America: O'Reilly, 2016. Citado na página 3.