Domain Application of Predictive Analytics Project Report for Loan Default Prediction

1st Tushar Patil
MSc in Data Analytics
National college of Ireland
Dublin, Ireland
x19199988@student.ncirl.ie

***Abstract*—In continuation of our project design proposal, we have carried out our selected technique for loan default prediction and documented the various aspects related to the same. This project document comprises of critical investigation of possible methods to address loan default prediction task, our selected methodology to carry out the prediction backed up by various related research in this field and the quantitative and qualitative insights from the proposed method. After reviewing various possible methods available to carry out this prediction, we have selected a logistic regression model in our project. We have evaluated the performance it using a performance matrix and provided some interesting insights while keeping the business requirements in mind.**

***Index Terms*—Loan default prediction, machine learning, Power BI, visualization, banking, risk analysis, credit management**

I. Introduction

Financial lending has been an important part of our system. Financial institutes and banks are crucial contributors to the financial market. Loans are one of the major sources for any lending firm or bank to acquire profit and maintain sustainable business growth. Where on one hand loans are the biggest income source for financial institutes, on the other hand, it carries various risks like loan default, frauds, etc. As a result of this more and more financial firms are exploring the use of machine learning for detection and prediction to tackle such risks. An efficient loan default prediction technique not only saves the firm from losses generated by loan defaults but also help to minimize the various expenses associated with loan management. In this project, we are focusing our study on the effective prediction of loan default using the Logistic regression technique. This method can be extended to other similar challenges present in the industry. [1] [2]

In this project, we have critically reviewed various latest work done in this field and selected the above-mentioned technique. We are making combined use of predictive analytics and supervised machine learning technique to build a business solution for the early prediction of potential loan default cases for a lending firm. Our methodology consists of various steps like data gathering, data cleaning, explanatory data analysis, feature engineering, model training-testing and evaluation of model performance to help us match the business requirements. We are evaluating the performance of the Logistic regression model with help of a confusion matrix and providing some interesting insights with the help of data visualization. We are using the open-source data available on the below-given URL:- https://www.kaggle.com/gauravduttakiit/loan-defaulter

Fig. 1. Lending Word Cloud

II. Project Objectives

As discussed in our system design proposal document we have successfully managed to address below given project goals while keeping business point of view in mind using predictive analytics and machine learning.

- Prediction of default loan accounts.
- Detecting underlying features apart from financial features which results in the loan default.
- Use of informative visualization to uncover patterns present in data.

III. Research on Related Work

Financial lending has been an important part of our system. Financial institutes and banks are crucial Fraud detection, credit risk management, default prediction using machine learning are well-discussed issues present in the financial market. There are several pieces of research and studies associated with the same. In this section, we would discuss various related work done in this field and our choice of approach for the loan default prediction task.

[14] has carried out a comparative study of various machine learning method to handle credit card fraud prediction task which closely resembles our task. Here they have critically analysed the performance of different machine learning models based on different performance parameters like sensitivity, accuracy, precision, etc. and arguments the Logistic regression model to be outperforming other methods. [11] also carried out a similar experiment combined with feature engineering methods like LDA and PCA for dimension reduction. Here also they have concluded the superiority of the logistic regression model over the random forest and KNN methods. [12] has critically analysed the importance of some features over others for default prediction in credit risk management. They have provided experimental proofs to back up the hypothesis that some features are more important than others in default prediction.

[13] has studied and provided the advantages of machine learning techniques over existing methods like the FICO model. Here they also highlight the effect of unbalanced data on the performance of machine learning methods. They have used an alternative method of using weighted models to handle class imbalance challenge and arguments their superiority over other methods. [16] has discussed the pros and cons of different resampling methods available to tackle class imbalance issue in credit risk management study. Here they are differentiating the importance of various performance parameters used to evaluate the model performance.

IV. Methodology and Implementation

As discussed in project design we are focusing our study to predict potential default loan cases for financial firms. To achieve the objective(s) of our study we are making use of CRISP-DM methodology as it has various advantages like flexibility, long term strategy, etc. CRISP-DM closely addresses the business objectives which is the major objective of our study. Fig- 2 depicts the steps involved in the CRISP-DM technique.

Fig- 3 shows different steps involved in our project. As mentioned above we are making use of publicly available data from lending institutes. Our selected data consists of numerous features related to customer application like borrower's demographic, financial info, customer's asset information, documentation information and a binary target variable. As we have the historical data present we are using the supervised machine learning method for our application. Also, our target variable is categorical hence we are using the Logistic Regression model in our task.

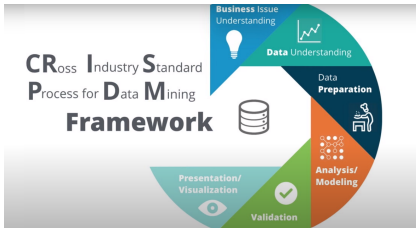

Fig. 2. CRISP-DM

Fig. 3. Project Flow

A. Data Cleaning

After acquiring the data we cleaned the data from null values. This is one of the important steps for better predictions in our analysis. We have removed the rows as well as columns having large null values(more than 50%). While removing these columns and rows we are not losing significant samples of data as the dataset consists of large sample values and features as well. By doing this we are also reducing our model processing time significantly.

B. Explanatory Data analysis

After getting rid of all the null values from the data we have carried out data analysis with the help of visualizations.

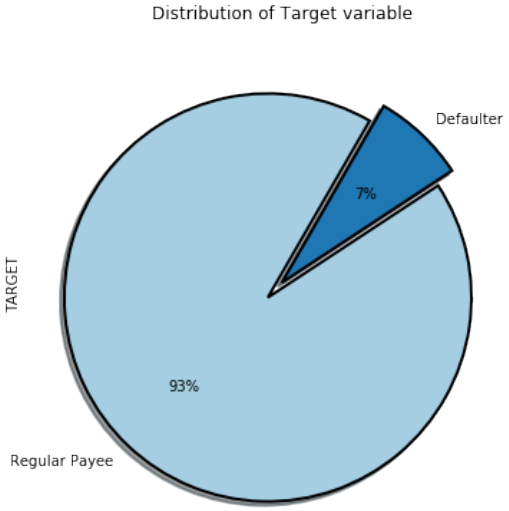

Fig. 4. Target feature Distribution

We are first analysing the distribution of our target variable. Figure-4 represents the percentage distribution of the same. It is clear from the figure that the data is critically unbalanced which is usually the common challenge associated with default prediction problems. we would be handling this issue by using an undersampling method which is discussed in the feature engineering section.

After this, we are examining the distribution and patterns present for different features present in the data. Figure-5 represents the gender ratio for loan applicants. It is evident from the graph that there are as many as twice female customers as compared to males.

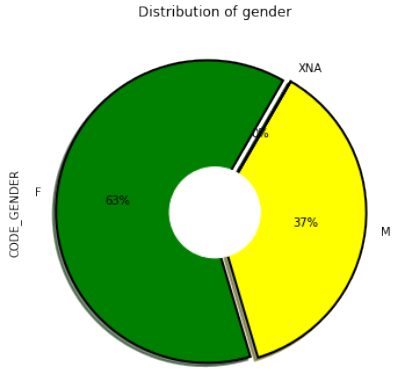

Fig. 5. Gender Distribution

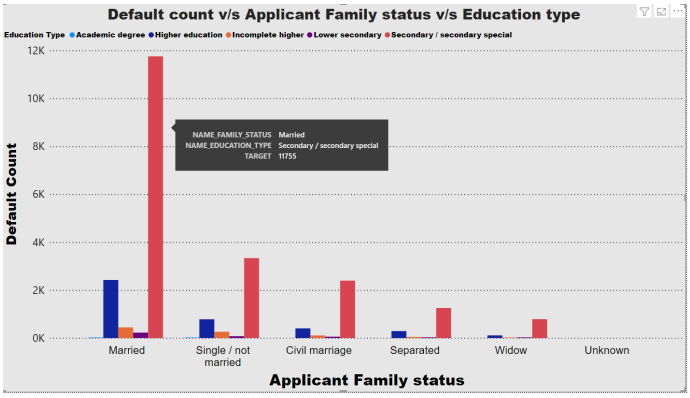

Fig. 6. Family status and Education level for Defaulters

To dive deeper into the analysis we have mapped the marital status and education level for the default loan cases. It is observed that married persons with education till higher secondary level are more likely to result into defaulters. Figure- 6 also explains that their people with mid-level education are more inclined towards loan defaulting as compared to higher educated and non educated people.

The applicant's occupation represents his financial stability. In Fig- 7 we are analysing the relationship between the occupation of applicants and their loan status. It is evident from the figure that having a stable work of applicant does

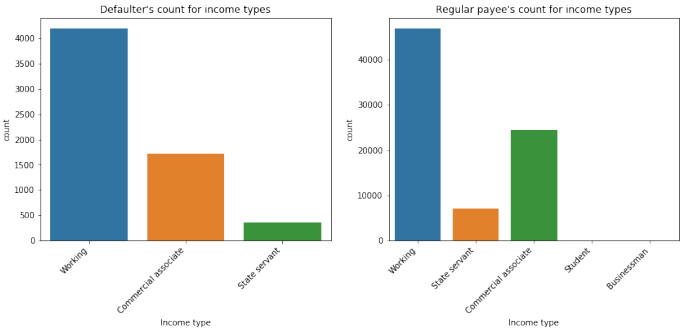

Fig. 7. Occupations for Both Applicant Types

not assure the non-defaulting nature of the applicant. The figure also reveals that people employed by commercial associations tend to be more regular with their due payments.

House, cars are used to analyse the financial sustainability of the applicant. Many banks and organization authorise the loans to only those who have some kind of equivalent assets as compared to their credit amount. In Figure-8 we are looking at that count in percentage form. It seems like almost 70% of the applicants have such properties which they could use for the mortgage if needed. Also in another figure, we are dividing this percentage using gender. It turns out that out of all property owners 64% applicants are females. The above analysis combined with Figure-5 confirms that despite having a higher percentage of asset ownership there are strong chances of females applicants turning into default customers.

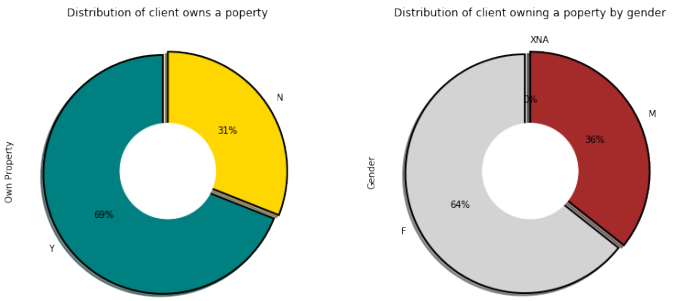

Fig. 8. Asset Distribution: Overall and Gender vise

To get a more clear idea about different features related to default applicants we have mapped various parameters like gender, contract type, car ownership, property ownership concerning applicant status(regular payee, defaulter) in Figure-9. It signifies that cash loans are observing a large number of cases in both scenarios.

We are also carrying out some statistical analysis on financial features like-AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE for both customer categories(regular payee, defaulter). Significantly, the mean income amount for defaulter customers is slightly higher than

Fig. 9. Cross Feature Distribution for both Applicant Types

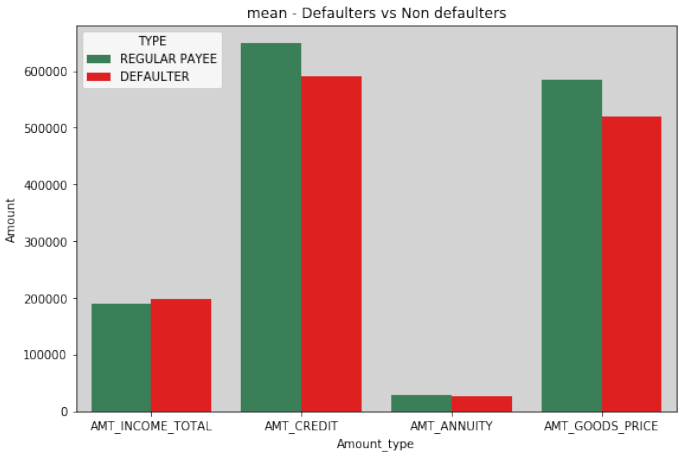

Fig. 10. Statistical Analysis for Different Amounts

regular payee customers. Also, credit amount does not affect loan defaulters as for both the categories mean credit amount is similar.

C. Feature Engineering

As discussed above our selected dataset contains a large number of features. Though all these columns contain important information, processing and handling more columns increases the chances of overfitting the model due to multicollinearity between the features. To avoid such poor performance we are making use of the principal component analysis(PCA) technique for dimension reduction. PCA makes use of the Eigen value for dimension reduction without losing the information carried by various features. In this project, we are carrying out PCA for features related to documents and credit amount required features due to the similarity between them. With the help of PCA, we have significantly reduced the feature count from the data without losing any information. [7] We also dropped some unrequired columns from the data after looking at the correlation matrix. Once we have finalized our data we have encoded the categorical features present in the data using label-encoder.

D. Model Training and Testing

Once the data is cleaned we are dividing it into training and testing datasets. We are using 80% data for training purpose and 20% data for testing the trained model. As we have detected earlier our target variable is poorly balanced in nature which results in poor performance of the logistic regression model. To avoid this problem we are using the undersampling method.

- **Random Undersampling:** Random undersampling is a sampling method that selects the samples from the majority class and randomly removes them from the dataset. This process is continued till the majority class percentage is achieved as per requirement. There are chances of losing important details while using this method if the given data is small in size. But in our case we have a large number of samples present in our data hence this is the ideal method for handling the class imbalance issue in our prediction task. [8]

- **Logistic Regression:** Logistic regression is a statistical machine learning method used to solve classification problems. It is an enhanced version of the linear regression model. It makes use of the Sigmoid function to categorize the target variable. It has many advantages like high flexibility, cost-effectiveness, high performance with large data, etc. which makes it a good and practical choice for our loan default prediction task. [9] [10]

To back up our methodology we are using two approaches in our project. First, we are training logistic regression(LR) on unbalanced data without using an undersampling method and evaluating the model performance on test data. In the second approach we are using the random undersampling method to balance our data. We are training LR on this balanced data and then evaluating its performance. The training time for the second approach was slightly shorter as the data is smaller in this case. Figure-11 shows the change in the distribution of the target variable for both approaches. We have tried different percentages of distribution(50:50, 75:25, 60:40) for the target variable and decided to use 50:50 proportion as it yields the best results.

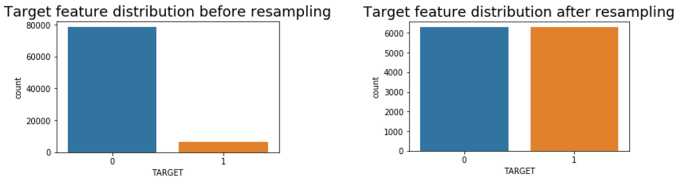

Fig. 11. Class Imbalance for Target Feature Before and After Undersampling

E. Model Evaluation and Results

We are evaluating the performance of the Logistic Regression model by using different performance parameters like

accuracy, precision, recall and F1-score. [4] [5] [6] Confusion matrix forms a basis for calculating this values. Figure- [**?**] introduces the terms associated with the confusion matrix and formulas for calculating different values.

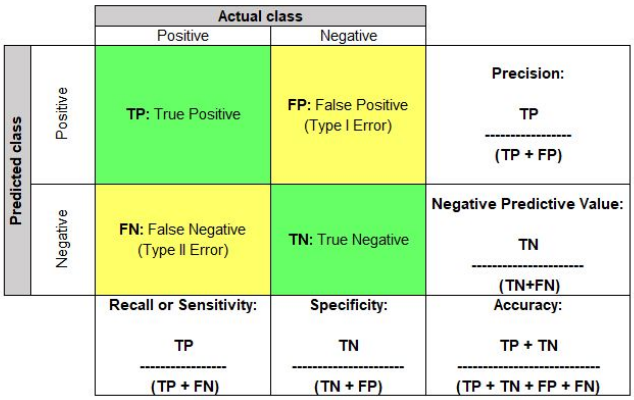

Fig. 12. Evaluation Matrix and Evaluation Parameters

Accuracy represents the overall performance of the predictive model for classification problems. Here when we are testing our model on unbalanced data we are getting an excellent accuracy of 92%. But due to the unbalanced nature of the data, the predictions are dominated by the majority class in this case. Due to this LR is predicting default cases as normal cases. This is evident in Figure-13. As we can see the count for False Negative values is very large. These are referred to as type-II errors.

In our prediction task, we are trying to achieve the best results for the prediction of possible default cases. With the above-given results, we can say that our model is miscalculating a large number of potential default loans which would cost a huge financial loss to the banks. As per the business objective of this task we are more interested in minimizing the type-II error as low as possible which would result in lesser financial loss. Here we cannot afford the higher number of false-negative count despite having high accuracy for prediction. Another parameter to check the performance of the model in the F1-score indicated how good the model is performing overall. Its value is between 0 to 1 and should be as close as possible to 1 for better results. It is confirmed that for the unbalanced data LR is performing poorly which can be verified from a lower f1-score in Figure-13.

As discussed above to overcome this poor performance we are making use of the undersampling method. By testing the LR on a model trained on balanced data we have achieved results shown in Figure-14. It is clearly visible that by balancing the data we have reduced the count for False Negative(Type-II) errors significantly. It is also evident from a much higher F1-score as compared to previous given results in Figure-13. While limiting the type-II errors in the prediction we are also slightly increasing the count for

```
BEFORE RESAMPLING
****************************************
Confusion matrix for Logistic regression model:
 [[15620     4]
 [ 1289     2]]

****************************************
Classification Report of Logistic regression model:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     15624
           1       0.33      0.00      0.00      1291

    accuracy                           0.92     16915
   macro avg       0.63      0.50      0.48     16915
weighted avg       0.88      0.92      0.89     16915

****************************************
```

Fig. 13. Classification Report: Without Undersampling

type-I errors where legitimate loan cases are also labelled as defaults. But as mentioned above in our business problem the cost of Type-II errors is far greater than Type-I errors hence there is a trade-off between these two values in our task. We have achieved an overall accuracy of 69% in our second approach.

```
AFTER RESAMPLING
****************************************
Confusion matrix for Logistic regression model:
 [[864 373]
 [412 862]]

****************************************
Classification Report of Logistic regression model:
              precision    recall  f1-score   support

           0       0.68      0.70      0.69      1237
           1       0.70      0.68      0.69      1274

    accuracy                           0.69      2511
   macro avg       0.69      0.69      0.69      2511
weighted avg       0.69      0.69      0.69      2511

****************************************
```

Fig. 14. Classification Report: With Undersampling

V. Conclusion

In this project report, we have successfully analysed various features resulting in loan defaults with the help of interesting visualizations. We have discussed explained the methodological steps involved in this prediction task and achieved optimum results while keeping the business objectives in the mind. While doing so we have faced the challenge of unbalanced data and came up with an undersampling method to overcome it. We have managed to attain the best possible results while keeping the trade-off between Type-II errors and overall accuracy.

REFERENCES

[1] https://www.grin.com/document/373126.

[2] Cornée, Simon, Soft Information and Default Prediction in Cooperative and Social Banks (June 17, 2014). Journal of Entrepreneurial and Organizational Diversity, Special Issue on Cooperative Banks, Vol. 3, No. 1, 2014, 89-109, Available at SSRN: https://ssrn.com/abstract=2450064

[3] https://mypeer.org.au/monitoring-evaluation/ethical-considerations/

[4] P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277458.

[5] https://www.scitepress.org/Papers/2018/68724/68724.pdf.

[6] https://arxiv.org/ftp/arxiv/papers/2002/2002.02011.pdf.

[7] https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db

[8] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman and A. Napolitano, "Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data," 2015 IEEE International Conference on Information Reuse and Integration, 2015, pp. 197-202, doi: 10.1109/IRI.2015.39.

[9] Li, Zhenchuan, Guanjun Liu, and Changjun Jiang. "Deep representation learning with full center loss for credit card fraud detection." IEEE Transactions on Computational Social Systems 7.2 (2020): 569-579.

[10] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[11] VidhiKhanduja and S. Juneja, "Defaulter Prediction for Assessment of Credit Risks using Machine Learning Algorithms," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1139-1144, doi: 10.1109/ICECA49313.2020.9297590.

[12] Sariannidis, Nikolaos, et al. "Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques." Annals of Operations Research 294.1 (2020): 715-739.

[13] Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction," 2020 International Conference on Computing and Data Science (CDS), 2020, pp. 212-218, doi: 10.1109/CDS49703.2020.00050.

[14] O. Adepoju, J. Wosowei, S. lawte and H. Jaiman, "Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978372.

[15] A. H. Alhazmi and N. Aljehane, "A Survey Of Credit Card Fraud Detection Use Machine Learning," 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020, pp. 1-6, doi: 10.1109/ICCIT-144147971.2020.9213809.

[16] Moscato, Vincenzo, Antonio Picariello, and Giancarlo Sperlí. "A benchmark of machine learning approaches for credit score prediction." Expert Systems with Applications 165 (2021): 113986.