Online Loan Application Prediction for a Housing Finance Company using Logistic Regression

Rohan Narayan Koli
MSc Data Analytics
National College of Ireland
x19224842@student.ncirl.ie

***Abstract*—For a Housing finance company, loan defaults remain the sole determining factor to make or break the business. Analyzing the borrowers at the earliest stages of loan application can minimize the defaulting risk to a larger extent and at the same time, fast-track the process to analyze thousands of such applications in parallel. In this research paper, we implement logistic regression as our predictive model based on historical records. The implemented model was able to achieve an accuracy of 84% using SMOTE data sampling technique. The research also found factors such as credit history, applicants income, co-applicants income, loan amount to be influential in determining the credit worthiness.**

***Index Terms*—Predictive Analytics, Machine Learning, Mortgage, Logistic Regression**

I. Introduction

The business model of a housing finance company relies on financial inter-mediation by raising capital and lending. The lending part of the company is highly exposed to risks and susceptible to frauds. In times where the COVID-19 pandemic has severely affected the world economy, the capital is sparse, there is cut-throat competition between peers and where the profit margins keep on decreasing, it is immensely essential to monitor the credit approvals and recoveries. The bottom line of the housing finance company is directly proportional on its ability to lend capital to non-defaulting clients and allow them to extend the credit as further as possible with a profitable interest rate.

For a housing finance company, to address the issue of defaults, predictive analytics is highly recommended. Predictive analytics was first implemented in 1950's by mail order industry to decide worthy and eligible applicants for credit. The decisions of credit worthiness were made by expert human underwriters by reviewing each application for credit worthiness [1]. Such systems can be implemented at the source where each customers application can be evaluated automatically online based on various features without human intervention. Predictive analytics encompasses various machine learning algorithms which are specific to each tasks. For this particular task we have implemented *Logistic Regression* as the predictive model to predict if a customer is likely to default or not based on the historic records.

In this research report, we implement a predictive model (logistic regression) to classify the online home loan applicants as approved or denied based on various features. The report is divided into four main sections starting with *Selection of the Predictive model*, where we compare and contrast the predictive the models previously preferred for the same task of loan approvals. Secondly, we implement the predictive model using the CRISP-DM (cross-industry standard process for Data Mining) methodology. Thirdly, we analyze the implemented model based on the evaluation metric and lastly conclude with the findings of the report.

II. Selection of the Predictive Model

A. Related Work

In the past, there has been various research conducted to discover the best model applicable for predicting the defaulting clients. We broadly categorize the researches into three categories namely, based on *decision trees*, based on *logistic regression* and based on *gradient boosting*.

1) Based on Decision Trees:

In [2], the author Anisa Ahmed implements machine learning algorithms to identify factors leading to high risk as well as assessing credit risk for small medium enterprises (SME). Three models, namely Naive Bayes, Random Forest and Decision trees were compared for classifying credit risk into seven levels (low to high risk). It was found that the decision tree model performed the best among other models with an accuracy of 99.99%. Also, the independent variables Real Estate principal amount and interest rate were instrumental in determining the grades of credit risk.

In another research [3], two predictive models were used, Decision Trees and Random Forest to make three predictions to asses if the loans will default or no, secondly, if the investment in loans will lead to profit generation and lastly, prediction of time of default. On analyzing the results, it was found that similar results were obtained by both the models, with Random forest excelling marginally against decisions trees. The evaluation metrics stated that Random Forest had a sensitivity of 89.63%, specificity 22.54% and precision of 88.70% compared to metrics of Decision trees which had sensitivity of 87.38%, specificity 17.20% and a precision of 88.75%.

Similarly, in another research [4], to minimize peer-to-peer lending risks and to evaluate the loans, five predictive models were implemented namely Logistic regression, Random forest, Decision trees, AdaBoost and Naive bayes. These models were tested on two scenarios, one where all the features were considered and secondly where only relevant features were

considered. Since, the data had class imbalance, synthetic minority over-sampling technique (SMOTE) was implemented to balance the data for better prediction. On evaluating the models, it was found that, in both the scenarios, decision trees performed the best with a consistent accuracy of 87%, precision of 81% and a recall of 96%.

2) Based on Logistic Regression:

In [5], for prediction of charged-off loans in a peer-to-peer lending business, the researcher made use of four predictive models namely, logistic regression, random forest, k-nearest neighbor and artificial neural networks. Notably, the logistic regression model with stochastic gradient descent(SGD) achieved superior accuracy of 71.14%, precision of 47.5% and a recall of 9.5% in classify the charged-off loans.

In another research to analyze factors contributing to loan defaults in a micro finance company (Grameen Bank) in Bangladesh, the researcher makes use of binary logistic regression to predict loan default by the borrower [6]. The model was successful in classifying 93.30% of the cases. It was also found that features like being a single borrower, higher repayment amount, lower interest rate were the contributing factors for loan defaults.

In [7], logistic regression is used to predict loan defaults for LendingClub which is a short-term micro loans network platform. The study also analyzes the factors affecting the default risk. The model acquired superior results with accuracy of 92.80% to predict the loan defaults and lastly, it was found that clients were more likely to default if they borrowed higher loan amounts.

In another research [8], the researcher implemented a model for micro-borrowers in Ghana to predict credit card defaults. Logistic regression was successfully implemented as the predictive model which fetched an accuracy of 76.1%. To predict the deafults, the researcher found eight influential factors namely- income level, gender, age, residential status, marital status, number of dependents, tenure and loan amount. Males had easier access to credit compared to females which could be due to cultural differences. Lastly, income status coupled with relative income (ratio of expenses over income) were significant in predicting defaults.

3) Based on Gradient Boosting: In a research [9], five models namely logistic regression, random forest, gradient boosting, neural networks and ensemble models were used to predict loan defaults. Additionally, three data sampling techniques such as SMOTE, ADASYN (Adaptive synthetic sampling) and cost sensitive learning techniques were used. It was found that using base model and class weights sampling technique, XGBoost algorithm had a superior performance compared to others with an accuracy of 70% and 68% respectively. The researcher further suggests implementing the model improves the ROI (Return on Investment) by 83%.

Similarly, in another research [10], XGboost gradient boosting algorithm was utilized to predict bank loan defaults. The algorithm performed with an accuracy of 79%, precision of 97% and a recall of 79%. Among all the features, the location and age of the customer were most important features in predicting the loan defaults.

B. Rationale for Model Selection

From the works of research mentioned above, we can imply that for predicting loan defaults, three algorithms namely logistic regression, decision trees and gradient boosting were extensively used. Though all the three algorithms performed well in predicting the defaults, we select *Logistic Regression* as our predictive model for the following reasons-

a) Prediction of Probability: In a Logistic regression model, we pass the dot product of weights and descriptive features through the logistic function (logit) which is given below:

$$\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

where, the left side of the equality represents the log likelihood of the event and the right side is the dot product of the features and weights [11]. Hence, instead of predicting the actual class, the model predicts well calibrated probabilities of classes (either defaulting or non-defaulting). Such probabilities are essential in decision making to predict the defaults prior and accordingly set a cap (threshold limit).

b) Simple yet Powerful and Fast: Logistic regression is very simple to implement,interpret yet it displays superior performance using lower computing power. Less computing power (faster prediction) equates to lower hardware and efficient data and energy utilization unlike algorithms like SVM, Decision trees and kNN.

Faster prediction results lead to processing and predicting hundreds of loan application records within a short amount of time (in seconds).

c) Data requirements and handling: Logistic regression works accurately given a complicated linearly separable data and a few records (less training data). Secondly, large datasets with millions of records can be handled with ease using the logistic regression.

d) Generality and flexibility: While training the model, logistic regression will not over-fit and maintain the right amount of bias and variance. Using the regularization (penalty functions) techniques such as lasso, ridge and elasticnet, the model can be converted into a more generalized model.

Hence, the above reasons clearly justify the usage of *logistic regression* for predicting if a customer will default on payments or not.

III. Implementation of the Predictive Model

For implementing the predictive model, we will implement the standard CRISP-DM methodology which provides a structured, robust and a flexible approach to solve business issues. Fig.1 depicts the CRISP-DM methodology.

A. Business Understanding

Our primary objective is to predict if the online applicant will default on payments in the future. Secondly, we intend to investigate how each input features are responsible for the outcome.

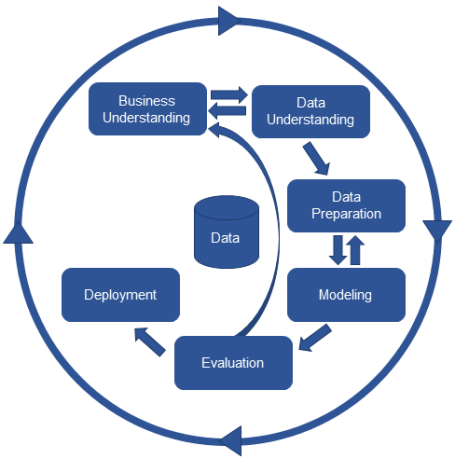


Fig. 1. CRISP-DM Methodology

B. Data Preparation

The data is collected from a practice problem on the website *AnalyticVidhya.com*[1]. The data contains 614 records and 12 feature columns along with 149 missing values as shown in table I.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)

TABLE I
DATA FEATURES

The missing values in the features *Gender* and *Married* were replaced by dominant classes, that is ”males” in case of gender and ”yes” in case of married. Next the number of *Dependents* were categorized into four categories(0, 1, 2, 3 or more). We created new levels in the features *Self-Employed* and *Credit History*. The missing values in *Loan Amount* were replaced by median. Lastly, the *Loan Amount Term* was categorized into five levels, namely between 0 and 120 months, between 121 to 180, between 181 to 300, between 301 to 360 and between 361 to 480 months. Next, the features ApplicantIncome, CoapplicantIncome and LoanAmount were standardized using min-max scaler. The data was split into test and train (65:35 ratio) using stratified split in the sklearn python library to maintain the class level ratios in each categorical features. Lastly, since our dataset is imbalanced, that is we have unequal levels of the dependent variable (422 Approved and 192 rejected), we use the SMOTE technique to create synthetic records for the minority class to improve predictions of logistic regression. The SMOTE technique creates synthetic records by oversampling the minority class as suggested in [12]. Lastly, we plot the co-relation matrix to analyze the

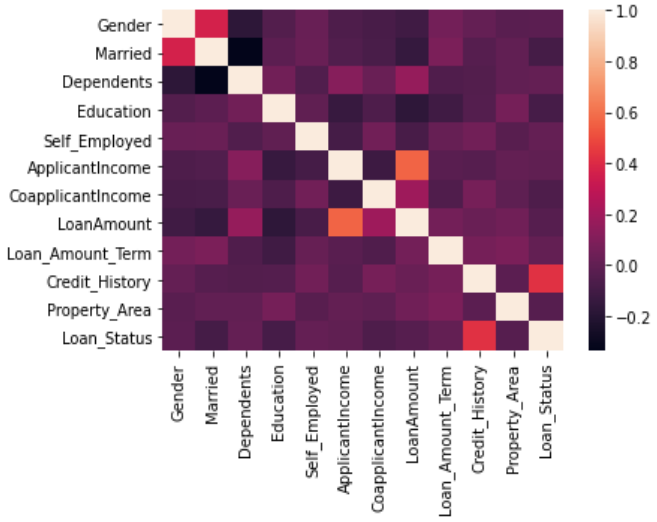


Fig. 2. Correlation Matrix

C. Modeling

After the initial processing of the data, we implement the Logistic regression model using the sklearn library in python. Since, in logistic regression we get predictions as probabilities, we convert the probabilities greater than 0.5 as ”yes” and less than 0.5 as ”no” for our dependent variable (Loan_Status). The modeling process is depicted in fig.3. The data from online applications is collected and stored in databases. Initially, past records which were analyzed by trained professionals are used as a base for training and tuning the logistic model. The data is split into training and test data and the model is evaluated in the next sections.

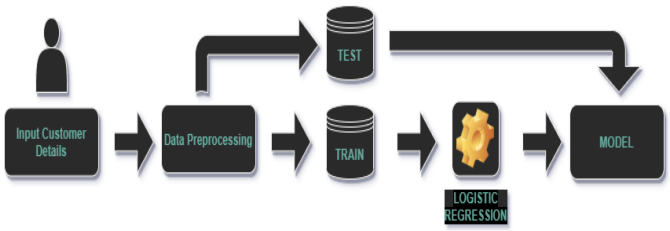


Fig. 3. Predictive Model

IV. RESULTS AND DISCUSSION

A. Model Evaluation

1) Confusion Matrix: The best way to evaluate a predictive model with classification task is to count the number of

[1]https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

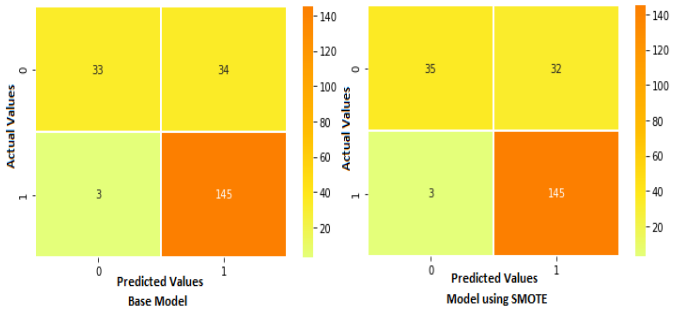


Fig. 4. Confusion Matrix

instances of class being correctly classified and misclassified in the confusion matrix as depicted in fig.4.

From the confusion matrices for the base model and the model with SMOTE, we get the counts fro true negatives (TN), true positives(TP), false negatives(FN), and false positives(FP). Based on the confusion matrices, we calculate the accuracy, sensitivity, specificity, F1-score for the model. They are defined by the following formulae-

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

On evaluation, we get the following results for both the models as depicted in the tableII.

Metrics	Base Model	Model using SMOTE
Accuracy	82.79%	83.72%
Precision	81.01%	81.92%
Sensitivity	97.97%	97.97%
F1-Score	0.887	0.892

TABLE II
EVALUATION METRIC

2) Receiver operating characterisitic curve (ROC): The ROC curve is created by plotting false positive rate (FPR) on x-axis against true positive rate(TPR) on y-axis at various decision thresholds as in fig.5. On inspecting the curve, an

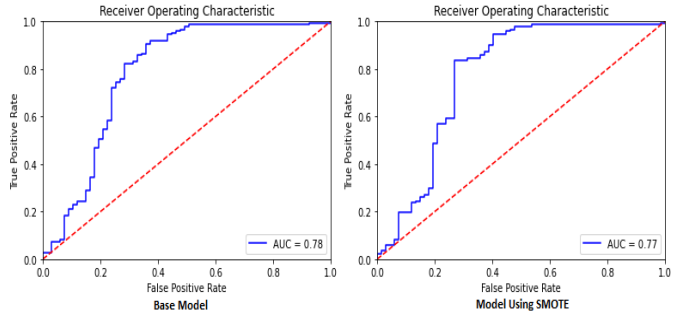


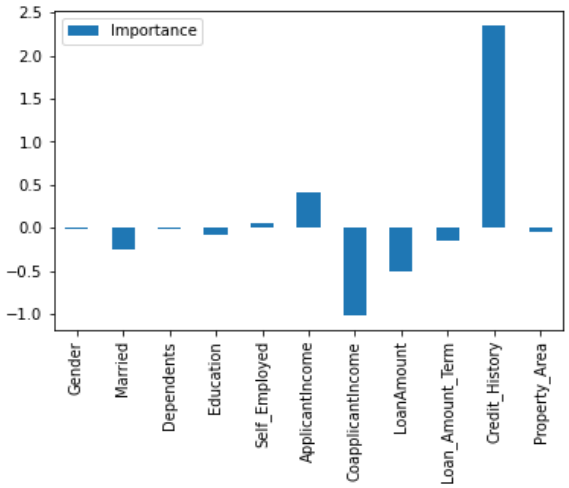
Fig. 5. AUC-ROC curve

AUC of nearly 77% is achieved with both the models.

B. Analysis and Interpretation

Overall, the model was able to perform with superior results with nearly 84% accuracy, 82% precision, 98% sensitivity and a F1-score of 0.88. Hence, the model can be deployed successfully for the housing finance company (HFC).

To analyze the model in-dept, we plot the feature importance matrix by raising the power of Euler number to each coefficient of the independent variables as seen in fig.6.



Features	Importance
Gender	-0.02
Married	-0.26
Dependents	-0.01
Education	-0.08
Self_Employed	0.06
ApplicantIncome	0.42
CoapplicantIncome	-1.01
LoanAmount	-0.51
Loan_Amount_Term	-0.14
Credit_History	2.34
Property_Area	-0.04

Fig. 6. Feature Importance

The credit history is the strongest feature in the loan dataset. By having a credit history, the applicants odds for his loan amount being accepted is raised by a factor of 2.34. The credit history is defined by our past credits which we might have accepted such as auto loans, personal loans, credit cards, education loans and so on. Each individual on accessing credit, they are rated in between 300-850 depending on various factors using the FICO scoring system (higher number suggesting an excellent record whereas lower number suggests a poor score). The scoring is dependent on various factors such as payment history, credit amount, tenure, type of loan and number of new credits. Hence, the credit history acts as a litmus test for loan approvals for the HFC.

Secondly, having higher co-applicants or co-borrowers income leads to more loan application refusals i.e for each unit increase in co-applicants income, the odds of loan approvals reduce by a factor of 1.01. Although, co-borrowers with a good

credit history and higher FICO score can help lower interest rates and increase loan eligibility, having higher co-applicants income in the first place can lead to more refusals.

Thirdly, an increase in loan amount leads to refusals by a factor of 0.51 which suggests application for higher loan amounts tend to get rejected. Higher loan amounts questions the borrower's ability to repay the loans within the fixed term. This could be due to the fact that the borrower might have low income, savings or investments and hence they wont afford to re-pay their installments in time.

A supporting factor to the above point is, a unit increase applicants income leads to increase in acceptance by a factor of 0.42. Although higher income leads to higher credit availability, most mortgage providers cap their credit amounts to 3.5 to 4 times the applicants gross income.

Next, being married reduces the odds ratio of loan approval by a marginal factor of 0.26. Ideally, the odds of loan acceptance are increased if both the partners are employed (dual income) and both of them have a good FICO score. On the contrary, if one of them is unemployed (more commitments or burden for one of the partners) or has a bad credit record may lead to refusal if the loan application is jointly applied. In the case of our dataset, we may assume the latter.

Features like gender, number of dependents, education level, being self employed, property area for which credit is sought have minuscule impact on the approval criteria of loan application. Pragmatically, these factors should not impact the decision process else they may raise ethical and legal concerns. The predictive models implemented should strictly be gender unbiased and equal opportunity should be given to both the gender.

Lastly, less loan term amount leads to lower acceptance of loan applications. The HFC's want their customer to be engaged with them for longer periods to profit [13]. Early loan payments leads to lack of interest returns to the HFC's which minimizes profit. Hence, loan applications with shorter term have high rejection odds.

V. Conclusion

In this research paper, we successfully implemented logistic regression as our predictive model to predict loan defaults for an HFC by analyzing online loan applications. We utilized the CRISP-DM industrial standard to implement our model. The model was able to predict loan defaults with superior accuracy of 84% using the SMOTE sampling technique. Further analysis of the model suggested, credit history, marriage, applicant income, co-applicant income, loan amount, loan tenure to be the most decisive factors in granting of mortgage.

The COVID-19 pandemic led to a lack of capital availability among many individuals and ultimately impacting many businesses. Among them, the mortgage and real estate businesses were highly impacted as buying and selling houses was the last thing on the minds of consumer leading to a slump in housing prices and interest rates. Therefore, when the business is already facing difficulty, risking essential capital as loans is can lead to devastating effects on the companies bottom-line. To analyze customers at their initial application stages and eliminate any risks in of utmost importance.

Secondly, due to cut-throat competition among peers, loosing opportunities is also not favorable to the business. Hence, implementation of predictive analytic systems is inevitable. Therefore, such model can improve efficiency for the business and minimize risks. Further, periodically evaluating the model and fine-tuning the model with additional records can make the model in a more generalized form, ultimately influencing the business bottom line.

References

[1] S. Finlay, *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*, ser. Business in the Digital Economy. Palgrave Macmillan UK, 2014. [Online]. Available: https://www.palgrave.com/gp/book/9781137379276

[2] A. N. Ahmed, "Credit-risk assessment of small business loans using naïve bayes, decision tree and random forest," p. 25, 2019. [Online]. Available: http://norma.ncirl.ie/4220/

[3] H. M. Owen, "A machine learning approach to predict loan default," p. 85, 2019. [Online]. Available: https://repository.lib.ncsu.edu/handle/1840.20/37034

[4] A. Ayantola, "Minimizing credit risk in peer-to-peer lending business using supervised machine learning techniques," p. 25. [Online]. Available: http://norma.ncirl.ie/4317/

[5] B. Bhardwaj, "Prediction of charged-off loans for p2p online banking using classification models and deep neural network," p. 22.

[6] M. Aslam, S. Kumar, and S. Sorooshian, "Predicting likelihood for loan default among bank borrowers," vol. 11, no. 1, p. 318. [Online]. Available: http://www.sciedupress.com/journal/index.php/ijfr/article/view/16405

[7] T. Deng, "Study of the prediction of micro-loan default based on logit model," in *2019 International Conference on Economic Management and Model Engineering (ICEMME)*. IEEE, pp. 260–264. [Online]. Available: https://ieeexplore.ieee.org/document/8988492/

[8] K. S. Ofori, E. Fianu, K. Omoregie, N. A. Odai, and F. Oduro-Gyimah, "Predicting credit default among micro borrowers in ghana," p. 10.

[9] L. Zhu, "Predictive modelling for loan defaults," p. 42, 2019. [Online]. Available: https://escholarship.org/uc/item/3rs9b3d6

[10] R. Odegua, "Predicting bank loan default with extreme gradient boosting," p. 5.

[11] J. Kelleher, B. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, ser. The MIT Press. MIT Press, 2015. [Online]. Available: https://mitpress.mit.edu/books/fundamentals-machine-learning-predictive-data-analytics-second-edition

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," p. 37, 2002.

[13] E. Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley, 2016. [Online]. Available: https://www.wiley.com/en-au/search?pq=%7Crelevance%7Cauthor%3AEric+Siegel