Design document for Predicting the loan status of online loan applications for a Housing Finance Company

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

***Abstract*—The housing finance company earns by receiving interest from the customers for the provided loans. However, the profits of the company solely depend on the repayment ability of the borrower. Hence there is a desperate need to employ a predictive model to validate the loan eligibility of the applicant. In this paper, we present a design for such predictive model by taking into account the business implications along. For employing the predictive model, machine learning algorithms such as decision trees, logistic regression, k nearest neighbour (k-NN) and support vector machines (svm) are considered.**

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

I. Introduction

The COVID-19 pandemic has had a severe impact on the financial services sector. The housing prices along with mortgage rates are at an all time low. While this would indicate a boon for home seekers, it has turned out to be a bane for housing finance companies. The pandemic has bought high credit risks along with refinancing opportunities impacting the mortgage industry. Hence, it has become an essential task to analyze the customers applying for housing loans using *predictive analysis (PA)*.

In [1], the author stresses that for a housing finance company, decreasing loss ratio is of utmost importance. Customers who are engaged in mortgages for a longer duration turn out to be profitable for the lenders. But, if a customer defaults on mortgage payment or if a customer repays the mortgage at earlier at once (pre-payment), leads to no interest amount collection (loss). Such risks are termed as "micro risks" which consolidate creating a major crisis. Such crisis can be averted using PA which includes machine learning to train predictive models on historic data.

A. Dataset Source

The dataset in consideration is taken from a hackathon from 'Analyticsvidhya.com' contains records of 614 customers who applied online for home loans with 13 features. There are 149 missing values spread across the 13 features. The data dictionary is shown in Table I. The source for the dataset is as follows:

https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#ProblemStatement

TABLE I
Data Dictionary

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	Co-applicant 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)

1) Independent Variable: There are total 11 independent variables, as shown in the Table I disregarding Loan_ID. Categorical variables include Gender, Married, Dependent, Education, Self-Employed, Credit_History and Property_Area whereas ApplicantIncome, CoapplicantIncome, LoanAmont, Loan_Amount_Term are continuous.

2) Dependent Variable: The Loan_status is the dependent variable which is categorical in nature with 2 levels. There are 422 records whose loan status is approved, where as 192 applicants were rejected.

B. Data Pre-processing

The dataset contains 149 missing values spread across 12 features which needs to be processed. The missing values in the features 'Gender', 'Married' and 'Dependents' were replaced by their respective mode. New categories were added in features 'Self_Employed' and 'Credit_History'. Features 'ApplicationIncome', CoapplicationIncome' and 'LoanAmount' were normalized using Min-Max scaling. The feature Loan_Amount_Term was categorized into 5 categories viz., below 120 months, between 120 and 180 months,between 180 and 300 months, between 300 and 360 months and lastly between 360 and 480 months. Finally, the categorical variables were coded numerically according to their respective levels.

II. Goals

The primary objective of the project is to create a model using predictive analysis (PA) by implementing machine learning to enables us to analyze and make decisions if the applicant is eligible for housing loan. The online application acceptance or rejection process can be automated using PA.

The secondary objective of the project is to analyze how each feature in the model, influences the decision variable. An analysis of each features will help us build characteristics of an individual (applicant), and finally help us get the predictive score for each record.

III. Ethical concerns

There are two major ethical issues that can arise if the predictive model was implemented.

1) Algorithmic Bias: A PA model is trained using machine learning on datasets containing historical data. In training a model, the algorithm can create biases when making decisions by assigning high weights to incorrect features. For example, a machine learning model may develop prejudice by giving more weights to the location of residency or the number of dependents the applicant has than to the credit history of the applicant to approve or reject loan application. Such algorithms can create unfair decision which are unethical. Proper scrutiny of the training data provided and back-testing of the algorithm needs to be carried before implementation.

2) Transparency: Most of the times, when a PA model is developed, the operation of algorithm as to why a particular decision is made in a particular case is unknown. The highly complicated models built act as a 'black-box' which ironically cannot be interpreted by the users of the algorithm as suggested by the author in [2]. For such explanations organizations come-up with *counterfactual* explanations which are unethical. These counterfactual arguments are made without the actual interpretation of how the algorithm actually works thereby, protecting the intellectual property.

3) Building Algorithms on Personal data: A PA model should not consider implementation of personal data such the gender or marital status in training the algorithms. Such algorithms should be treated illegal on ethical grounds.

IV. Business Value

Employing a predictive analytic model has many fold advantages in predicting if the online loan application should be accepted or rejected, some of which are discussed below.

A. Minimize human errors

"Probability of human error is considerably higher than that of machine error." - Kenneth Appel

Humans are prone to errors and biases. The judgement of the loan application grants cannot be solely entrusted on employees as they are susceptible to make mistakes. Errors in such judgments can cause huge losses to the company if left unchecked. Errors have a potential to directly impact the bottom line of a company. Huge amounts of loans if granted to wrong applicants can cause havoc to the housing finance company or may even have a serious industry-wide impact. The 2008 economic crisis was fueled by NINJA loans (No Income, No Job, and No Assets Loan) causing the housing market bubble to burst. Hence, proper modelling of the PA in loan application systems can be advantageous by minimizing the human element for eligibility validation.

B. Cost saving

A PA model validating the loan application process can save employee overhead cost for the company. The company has to hire few employees to handle the validation process. Most of the applications will be filtered through the PA model and the employees will have less applications to validate. Hence, the PA model can act as gate-keepers to the business process. Having less employees, the company can save on infrastructure as well as overhead costs (electricity, work stations, etc.) leading to better cost optimization.

C. Time saving

A PA system can process hundreds and thousands of applications with a shortest amount of time with highest accuracy and precision as compared to humans. In today's fast moving world, it is necessary to save time which is highly critical resource leading to customer satisfaction. Fast loan processing leads to more incoming business opportunities ad ultimately maximizing profits.

D. Robust System

The PA models have the ability to be versatile and adaptable. Thought the models rely on historical data, they can be trained on the changing trends of the market. The working of PA models are easily interpreted and they refrain from acting as black-boxes. Hence, the employees and managers of the company can take prompt decisions and make necessary modifications to the PA models in order to improve its efficiency. They act as transparent models, and help operate within the legal system boundaries.

E. Competitive edge

Now-a-days as the business learns about the importance of predictive analysis (PA), every organization incorporates the same for a competitive edge. Businesses demand being a step ahead of the competition for survival and ultimately profit generation. The predictive model helps in decreasing risks. It learns from the past data which has a record where the previous system failed and where the previous system succeeded. Each of these instances lead to learning of better and more precise algorithms (self-learning algorithms). Hence PA models provide competitive edge by taking into account customer satisfaction.

V. Visualizations

We perform the exploratory data analysis using PowerBI and python. On analyzing the data, it is found that applicants who have qualification as "Graduate" have overall applied for

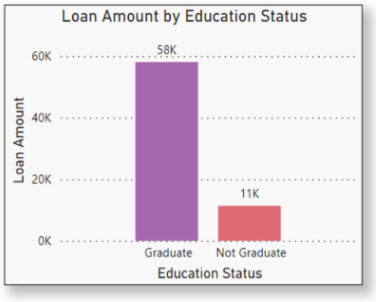

Fig. 1. Loan Amount by Education

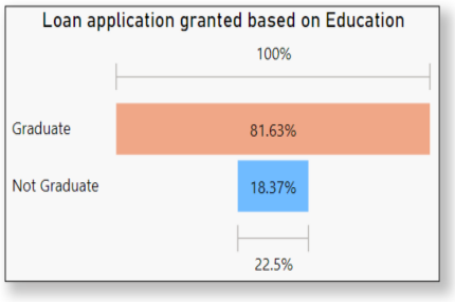

Fig. 2. Loan application grants based on Education

amount of loan compared to the non-graduates as depicted in Fig.1.

Further, from the dataset, it is found that graduates have a higher probability for housing loan acceptance as shown in Fig.2

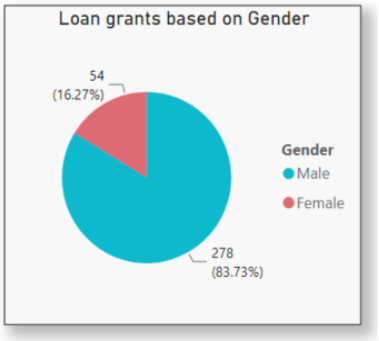

Fig. 3. Loan application based on Gender

Based on gender, it is found that, Males have higher likelihood of loan acceptance as in Fig.3. But this can also be due to the fact that, Males have applied in more numbers than females or due to the imbalance in the dataset.

Fig.4 illustrates that highest number of applicants opted for a loan term of 30 years whereas applicants who opted for 15 years of loan term were second highest.

The home loan approval probability was 68% for the people who were married as show in Fig.5.

The applicants who were not self-employed were granted more number of housing loans as in Fig.6.

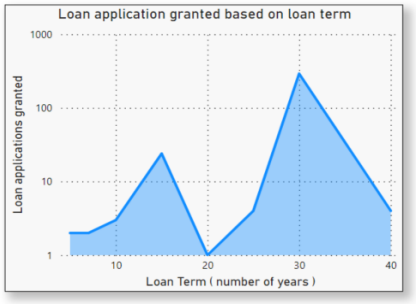

Fig. 4. Loan application based on loan duration

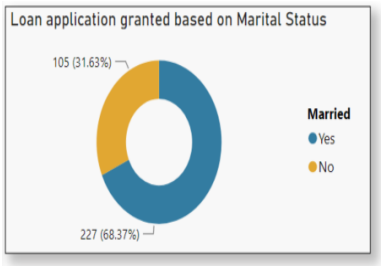

Fig. 5. Loan grants based on Marital Status

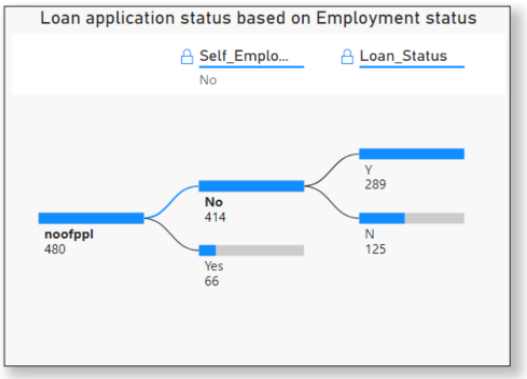

Fig. 6. Loan approval based on Employment

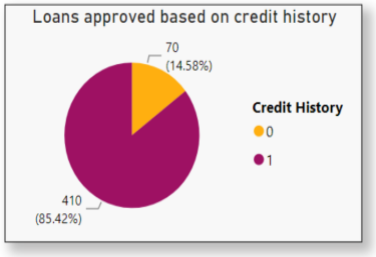

Fig. 7. Loan approval based on credit history

Fig.7 illustrates that, applicants who had a prior credit history were more likely to be accepted for home loan applications.

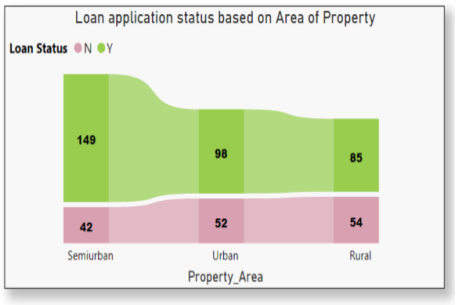

Fig. 8. Loan grants based on Property Area

From Fig.8 it is evident that semi-urban population were granted more housing loans, followed by urban and rural population of applicants.

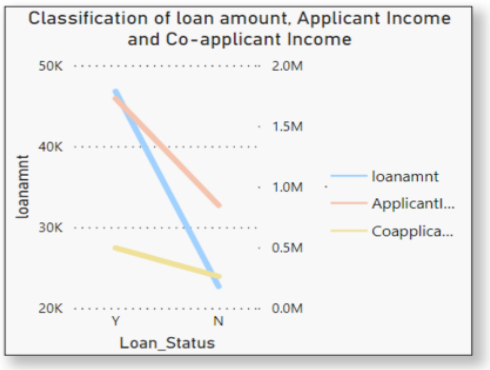

Fig. 9. Loan application based on Income

The population of applicants with higher income, along with higher income of co-applicants had a better probability of housing loan being granted.

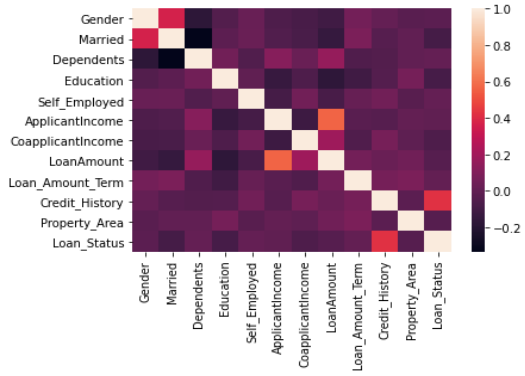

Fig. 10. Correlation Matrix

Lastly, Fig.10 illustrates the correlation heat-map matrix among the features with respect to the outcome variable. On analyzing the heat-map, it is found that the outcome variable (loan_status) is more influenced by or has a higher correlation to credit history of the applicant.

VI. Applicable Techniques

Since we are predicting if the home loan online application by the applicant is to be approved or not, it is a binary classification problem. There are four major machine learning algorithms which help in solving the binary classification problem, namely, Decision trees, Logistic Regression, k-nearest neighbour(k-NN) and Support Vector Machines (SVM).

A. Decision Trees

A decision tree algorithm models the relationship between features using a tree structure. The tree beings at the trunk (root node) and splits into narrower branches of decisions (decision nodes) as it propagates into a final predicted class (leaf node). Fig.1 illustrates a decision tree. The decision tree algorithm outputs a human-readable structure which is best suited for classification mechanism and displays transparency. Hence, the results about the business practices or legal aspects can be shared with others [3]. Decision trees utilize a divide and conquer approach by splitting the data into subsets repeatedly into smaller subsets based on heuristics (recursive partitioning). In [4] , the author compared five machine learning algorithms for predicting credit worthiness and found Decision Tree to perform the best classification task with 87% accuracy and 96% recall.

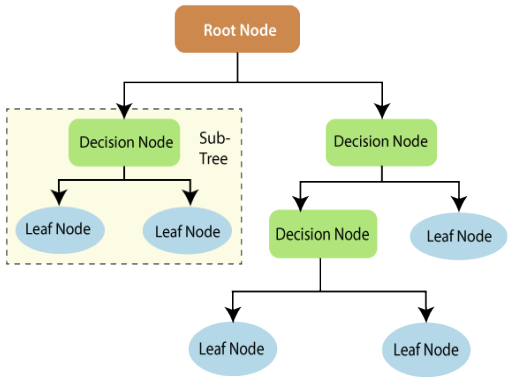

Fig. 11. Decision Trees

The advantages of decision tree algorithm are that they act as an all-purpose classifier, can handle numeric, nominal data along with missing values and they are interpreted easily. The disadvantages include biasing towards splits with more levels, over-fitting or under-fitting and some relationships are not modelled properly due to axis parallel-splits.

B. Logistic Regression

In [5], it was found that Logistic regression with SGD training got highest validation accuracy of 71% amongst three others. Similarly, in [6], logistic regression scored the highest area under curve (AUC) of 70% while predicting loan repayment. Logistic regression succeeded by overcoming the

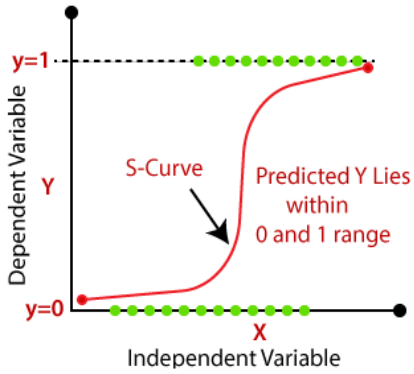

Fig. 12. Logistic Regression

linear regression drawbacks such as not producing probabilities between the range of 0 and 1 and assuming errors are statistically independent and normally distributed [7]. Hence, we utilize the logistic function (logit) expressed as -

$$M_d = \frac{1}{1 + e^{-w.d}}$$

Using the logistic function we get a 'sigmoid' curve which helps to model probabilities between 0 and 1 as illustrated in Fig.2. Logistic regression takes into account the natural log of odds ratio and models the probabilities of the dichotomous target variable.

C. k-Nearest Neighbour (kNN)

K-NN algorithm is used mainly for classification and it assumes the basic principle that similar things appear near each other in proximity. It calculates the number of points (given by 'k') on a graph with least distance (Euclidean distance) between them.Fig.3 depicts the working of k-NN algorithm.

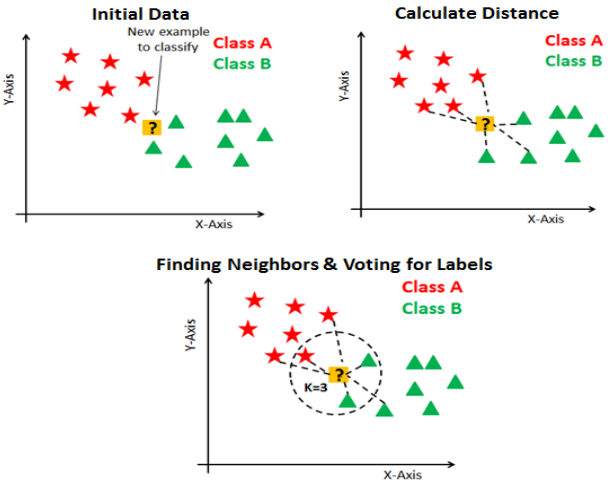

Fig. 13. k-Nearest Neighbour

In, [5], KNN was a competitive algorithm in prediction of charger-off loans with almost 70% validation accuracy. K-NN algorithm is easy to implement, fast in calculation and it does not make prior assumptions about the dataset. On the other hand, while implementing, we need to calculate the optimal value of the nearest neighbour 'K' every time.

D. Support Vector Machines (SVM)

SVM creates surfaces (hyperplane) which acts as a boundary to separate data based on feature values into similar groups. SVMs can be used for classification as well as numeric predictions, but binary classification is easily interpreted. The

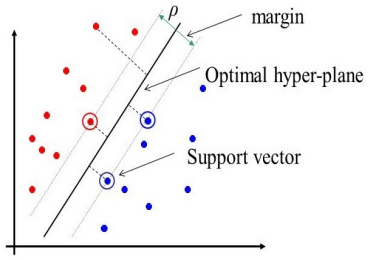

Fig. 14. Support Vector Machines

algorithm searches for maximum margin hyperplane (MMH) that creates the most optimal separation between the two classes relying on vector geometry [3]. In [8], SVM algorithm was compared with logistic regression for predicting P2P loan acceptance and default.

VII. Conclusion

This paper was aimed at providing a design document to implement a predictive model to validate the eligibility of online loan application. The business implications were discussed by taking into account the ethical concerns that would arise after implementation of the model. Lastly, various machine learning algorithms were discussed which would help construct the PA model.

The COVID-19 pandemic has bought the mortgage industry to a standstill with all-time low housing costs and lowest interest rates. In these trying times, survival of the organization is difficult and customer retention is crucial. Implementing a predictive model can act as a distinguishing factor for organizations giving them the supreme competitive edge over their competitions.

References

[1] E. Siegel, *Predictive analytics : the power to predict who will click, buy, lie, or die*. Hoboken, New Jersey Wiley, 2016. [Online]. Available: https://www.wiley.com/en-us/9781119153658

[2] A. Katwala, "How to make algorithms fair when you don't know what they're doing," WIRED UK, 12 2018. [Online]. Available: https://www.wired.co.uk/article/ai-bias-black-box-sandra-wachter

[3] B. Lantz, *Machine Learning with R*, 2nd ed. Packt Publishing, 2015. [Online]. Available: https://dl.acm.org/doi/book/10.5555/2876101

[4] A. Ayantola, "Minimizing credit risk in peer-to-peer lending business using supervised machine learning techniques," Master's thesis, Dublin, National College of Ireland, 2020. [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," Master's thesis, Dublin, National College of Ireland, 2020. [Online]. Available: http://norma.ncirl.ie/4433/

[6] C. Han, “Loan repayment prediction using machine learning algorithms,” Escholarship.org, 2019. [Online]. Available: https://escholarship.org/uc/item/9cc4t85b#author

[7] I. Witten, E. Frank, M. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2016. [Online]. Available: https://books.google.co.in/books?id=1SylCgAAQBAJ

[8] J. D. Turiel and T. Aste, “P2p loan acceptance and default prediction with artificial intelligence,” 2019. [Online]. Available: https://arxiv.org/abs/1907.01800