**Term Project**

**BSAN 775 – Introduction to Business Analytics**

**Fall 2022**

**(Class Meeting Day and Time: Tuesday, 7:05 PM)**

**Title:**

**Predicting the Creditworthiness of a Prospective Borrower**

**Team-9**

Sharat Chandra Vupparaboina

(A549F327)

Krishna Reddy Patlolla

(A637C336)

Yogananda Theeguru

(Q584C295)

**Table of contents:**

1. **Introduction………………………………………………….**  **3**
   1. **Background……………………………………………. 3**
   2. **Motivation……………………………………………… 3**
2. **Problem Statement…………………………………………. 4**

**2.1. Decision Time…………………………………………………………. 4**

**2.2. Bad Credit/No Credit history…………………………. 4**

**2.3. Customer Migration & Revenue loss…………………………………. 4**

1. **Objective…………………………………………. 5**
2. **Literature review…………………………………………… 5**
3. **Methodology……….………………………………………... 7**

**5.1 Data Description……………………………………. 7**

**5.2. Data Cleaning………………………………………. 7**

**5.3. Data visualization using Tableau………………………………. 8**

1. **Results & Discussion ………………………………………………………... 10**

**6.1. Logistic Regression…………. 10**

**6.2. Decision Tree analysis…………………………………... 12**

**7. Conclusion………………………… 14**

**8. References……………………………………………………… 14**

**1.Introduction:**

In this study, a publicly available dataset (<https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior>) is analysed to build a prediction model that can predict the probability of loan default by a borrower. The dataset is composed of several attributes of customers who have taken loans from financial institutions in India.

**1.1 Background:**

Loan lending is an essential part of several operations that take place in the financial industry. Lending loans has been a part of people’s financials from ages. Even when the banks aren’t around. The usual traditional way we often come across is a financial corporation or a bank lends out loan to a person or an organization who requests for it. A specific amount will be lent with a fixed amount of interest for a certain period of time. The lender will get his amount back along with the interest in instalments as agreed by both the parties during the transaction earlier.

Peer to peer is another form of lending where instead of an organisation directly lending money to a person, a person directly lends amount to another person through a lending platform or a lending organisation. This is called Peer-to-peer (P2P) lending. The organization is responsible for verifying the eligibility and creditworthiness of the borrower. Peer-to-peer (P2P) lending systems have much lower operating costs than traditional financial services. They provide a quick turnaround time for loan approvals.

**1.2. Motivation:**

According to Industry arc Research on P2P lending in India, the Indian P2P lending Market size is expected to grow at CAGR of 21.6% to reach $10.5 billion by 2026.

As discussed above, the P2P lending market is growing at rapid pace in India to take over a considerable pie of the circle in total lending market, with such huge market comes in the inevitable risk of customers not repaying the loans in time or completely defaulting on their loan.

The purpose of this project is to analyse the patterns between borrowers characteristics (independent variables) and the probable of risk of loan default (target variable) using the selected dataset.

**Keyword:** lender, borrower, online lending platform, lending organisation, creditworthiness, logistic regression, decision trees, visualization, accuracy, sensitivity.

**2. Problem Statement:**

Peer to Peer lending is growing more popular in today’s world as we lean towards digital approach for all types of activities and wants, and it comes with its own set of problems.

**2.1 Decision Time:**

When a borrower approaches a typical financial institution such as Bank, the decision time taken by the bank whether to approve the loan or not typically ranges from a few days to even months in some cases in India. This is a major problem for some borrowers who are in need of resources at the earliest, and P2P lending platforms offer a quick decision on approval/rejection of the loan request.

**2.2 Bad Credit/No Credit history**:

In most financial institutions, the first step in starting the loan application process is to check for the credit history of the customer and if the customer does not have a good credit history/ No credit history then it becomes very difficult for the customer to get a loan approved. There are millions of people in unorganized sector in India who may fall into such category and P2P lending is a great opportunity for such customers who do not stand a chance with Banks.

**2.3 Customer Migration & Revenue Loss:**

Imagine a Lending platform that does not have any type of credit rating in place, in such scenario lenders have no judgement on prospective borrowers before they can lend out their money. When a lender is faced with a borrower not repaying their loan, they lose trust on the platform. Moreover there are no investor protection regulations in place as far as P2P lending in concerned in India. This leads to lenders opting out of the platform to another trustworthy platform leading to revenue loss for the platform as the platform thrives on commission received from each loan approval.

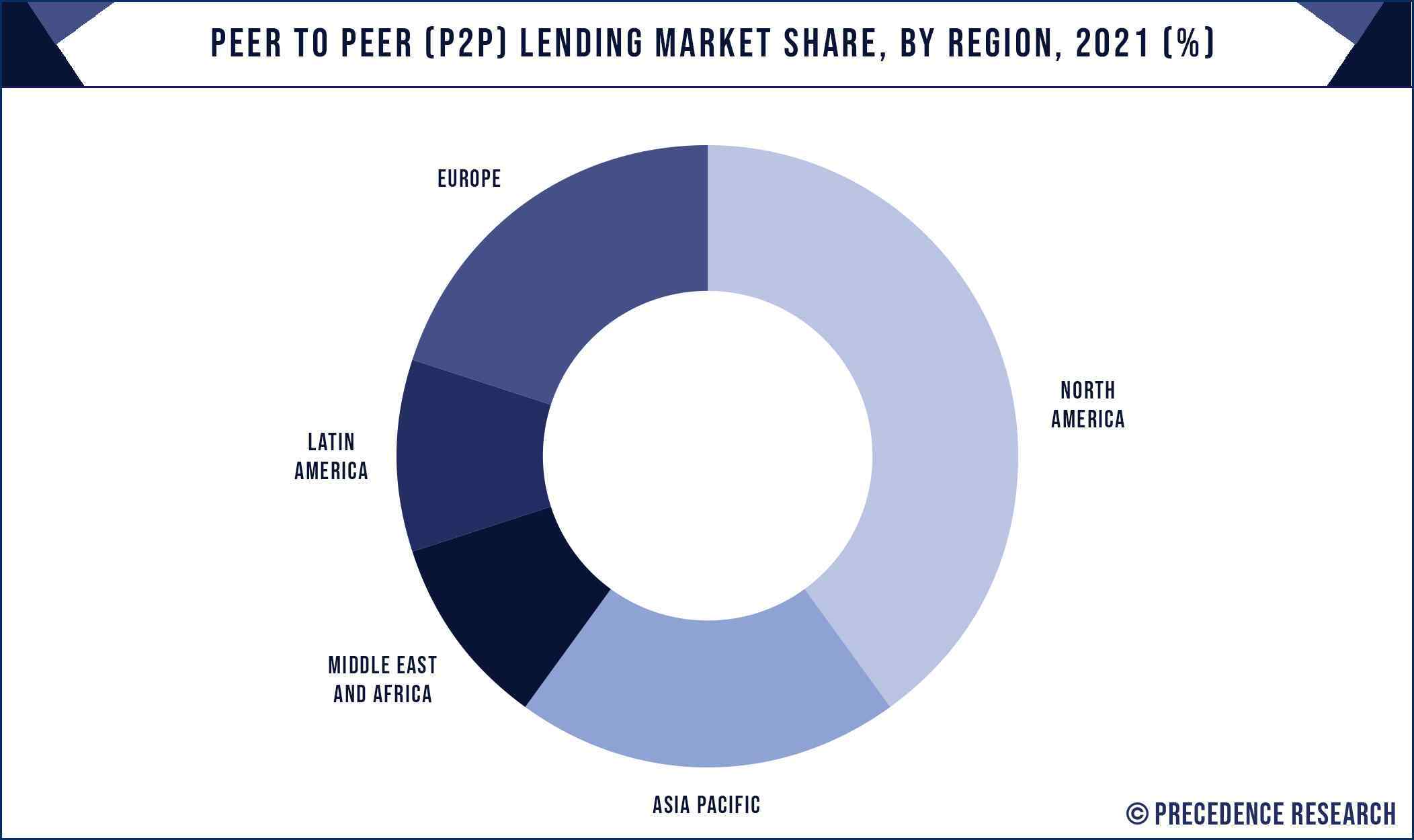
In this study we will analyse if the profession and experience of the customers have a significant impact on the creditworthiness of a new borrower.

**3. Objective:**

The main objective of this study is to assign a credit score to all customers who sign up on the lending platform as a borrower. Other objectives include finding the most significant attribute among the independent variables in predicting the loan default, to get a decent classification accuracy using decision trees and logistic regression. If implemented well this results in customer retention and revenue increase for the lending platform.

**4. Literature review:**

The global peer to peer (P2P) lending market size was valued at US$ 83.79 billion in 2021 and it is expected to hit over US$ 705.81 billion by 2030 with a registered CAGR of 26.7% from 2022 to 2030. The following image depicts the market share of regions in the P2P lending market across the world in the year 2021.



**Fig 1. Peer to Peer lending market Share**

In a research paper published by researchers from a Chinese university, to estimate the default risk of P2P borrowers in China, they suggested a multi-round ensemble learning model using heterogeneous ensemble frameworks made up of XGBoost, DNN, and LR individual learners.

It was discussed in a study “Internet Based Social Lending: Past, Present and Future" by "M. Hulme and C. Wrigh" that need for online lending has emerged from lack of funding to new customers with no credit history or bad credit history.

It was also opined in a journal "bank misconduct and online lending" by "Bertsch, Christoph & Hull, Yingjie (2020)" that banks are not serving lower end of customers as banks have high maintenance costs t and reputation to uphold.

Majority of studies revealed that total years of experience of a customer has negative effect on probability of default, this negative relationship between experience and loan default will be studied as one of the hypotheses in this study. It also to note that most of the studies conducted were based on the data obtained from USA and European nations, whereas this study is based on the data from India.

**5. Methodology:**

**5.1. Data description:**

The dataset has a total of 250,000 records with 13 attributes as follows:

**5.1.1 Predictor Variables:**

Income, Age, Experience, Married/Single, House\_Ownership, Car\_Ownership, Profession, City, State, Current\_House\_Years, Current\_Job\_Years

**5.1.2 Dependent Variable**

Risk\_Flag is the Target Variable which is a binary variable.

**5.1.3. Unused variables**

The “id” variable was unused in the model building as it doesn’t offer any statistical value to the model.

**5.2 Data Cleaning:**

**5.2.1. Unwanted Characters:**

There were no missing values in the data but upon observation we have noticed that there are some random characters like “#, $, \*” at the end of city names in City attribute, we have cleaned such unwanted characters using pandas in Jupiter Notebook.

**5.2.2. Handling imbalance**

Using Tableau, we visualized no of positive classes vs negative classes in the target variable, we have observed that there is high imbalance as the ratio of positive to negative class was 7:1. We used Random oversampling method in python to oversample the data to achieve an acceptable 1.6:1 ratio.

**5.3** **Data visualization using Tableau:**

Chart, pie chart

Description automatically generated

Fig 1. A pie chart showing the classes of our dependent variable.

Graphical user interface

Description automatically generated

Fig 2. A heat map of all the variables which depicts the correlation among the variables.

From the above heat map, we can see that the variable ‘Experience’ has the highest significance with our dependent variable ‘Risk Flag’ followed by ‘House ownership’ and ‘Married/Single.’ Following are the visualizations of these variables against the dependent variable.

Chart, bar chart

Description automatically generated

Fig 3. A stacked bar chart of experience variable against the dependent variable.

Chart

Description automatically generated with medium confidence

Fig 4. A horizontal bar chart of independent variable ‘House ownership’ against the dependent variable.

Chart, bar chart

Description automatically generated

Fig 5. A horizontal bar chart of independent variable ‘Married/Single’ against the dependent variable.

**6. Results & Discussion:**

* 1. **LOGISTIC REGRESSION USING SPSS:**

1. **WITH CUTOFF VALUE AT: 0.5Table

   Description automatically generated**

Fig 6. Table showing different coefficients of all the independent variables.

Table

Description automatically generated Table

Description automatically generated

Fig 7. Model summary of Logistic Regression Fig 8. Hosmer and Lemeshow test

Table

Description automatically generated

Fig 9. Classification table of logistic regression with accuracy when cut off values 0.50.

It is evident from the above results that logistic regression at cut off value of 0.5 is a bad fit for this model as Nagelkerke R-square is 0.011, which is extremely poor in predicting the variation in Target variable. Also, the model’s overall accuracy is around 62% with a sensitivity of around 99% and specificity of around 2% respectively. The model is not able to predict 0s in the dataset, therefore Binary Logistic Regression is not a good fit for this dataset.

1. **WITH CUTOFF VALUE AT: 0.75**

**Table

Description automatically generated**

Fig 10. Table showing different coefficients of all the independent variables.

Table

Description automatically generatedTable

Description automatically generated

Fig 11. Model summary of Logistic Regression Fig 12. Hosmer and Lemeshow test

Table

Description automatically generated

Fig 13. Classification table of logistic regression with accuracy when cut off values 0.50.

From the above classification table, we can observe when we change cut off value from 0.5 to 0.75 the model performs even more badly as the classification accuracy has come down from an already bad 62% to 39%. Also the sensitivity of the model is “0” and specificity is “100”, meaning that model is not able to predict true positives at all.

**6.2 Decision trees:**

With the poor performance of the Logistic Regression, in order to get a better predicting model, we tried Decision Trees, and the results were improved by a really good margin. We chose CHAID (Chi square Automatic Interaction Detection) as the growing method. This method detects the chi-square of the independent variables and picks the variable as the root node with the highest chi-square. We also applied cross validation with 10 folds. CHAID algorithm can have a maximum depth of 3. From the above classification matrix, we can derive that the model has specificity of 72.5% and a sensitivity of 94.3% and an accuracy of 86%. Minimum cases in parent node and in child node are chosen as 50 and 25, respectively. We got 948 nodes in the final tree.

Table

Description automatically generated

Fig 14. Classification table of Decision Trees.

If we consider the hierarchy of the variables we got in the decision trees output, “experience” is the root node followed by “profession” and “city” as the maximum depth is three. This also tells us that these independent variables play a major role in predicting the dependent variable class. If we can further explore much more deeply with more depth, we can have more variables.

In order to achieve our objective of finding out probable loan defaulters i.e., trying to identify the risky customers, ‘1’ in risk flag, we have to come up with different combinations of the aforementioned 3 most statistically significant attributes with different ranges. As we have too many nodes there are multiple leaf nodes which has a probability of ‘1’ i.e., risk of defaulting on loan is higher. Therefore, for each leaf the hierarchy is different. The following are some of the combinations with highest probability of ‘1’.



Fig 15. Table containing some of the derivations from the decision trees output.

From the above figure we can infer that if a prospective borrower matches with the specified values in those 3 attributes, then we can conclude that there is high probability of loan default happening. The above figure (Fig 15) only describes some of such combinations and there are many other such combinations that we can derive from the decision tree and use in arriving at a decision in giving a credit score to the borrower on the platform.

**Conclusion:**

Our hypothesis of experience being the most significant variable in the model has been validated with it being the parent node in the decision tree. Lenders can now see the credit scores given to the borrowers based on the decision tree results and make an informed decision.

**References:**

•Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 2009, China.

•Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483-3488.

Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on July 20, 2018