AUTOMOBILE LOAN DEFAULT PREDICTION
Group Name: Speed Racers

DATA REPORT

Group Members

1. Brenda Chepkoech
2. Kevin Kilonzo
3. Ted Kimani
4. Mercy Cherotich
5. Kelvin Kilel

1. Business Understanding

Business Overview

A non-banking financial institution (NBFI) or non-bank financial company (NBFC) is a Financial Institution that does not have a full banking license or is not supervised by a national or international banking regulatory agency. NBFC facilitates bank-related financial services, such as investment, risk pooling, contractual savings, and market brokering.

NBFIs offer loans to its customers to bridge various financial needs that include car loans. They compete with various other financial institutions to make this lucrative business opportunity profitable. They can give full car financing or a certain percentage of the total request depending on the various attributes of the customer.

NBFIs lack statutory recovery tools thus they tend to struggle to make profits due to loan defaults by the client hence the need to look into the various factors that affect their repayment abilities.

Business Objectives

To look into factors that affect client repayment ability and come up with a model that predicts whether a client will default or not

Business Success Criteria

Finding the factors that affect client repayment ability and having a model that accurately predicts

Whether a client defaults or not.

Assessing The Situation

1. **Resources inventory**

a. The dataset
   The link to our dataset is
b. Software
   Github, Tableau, python

2. Assumptions

The data is correct and up to date

3. Constraints

The data contained missing values hence needed to be cleaned

Data Mining Goals

1. To determine if the gender of the applicant affects the default rate
2. To find out how the occupation of the applicant affects the default rate
3. To determine if the age of the applicant affects the default rate
4. To determine if having different loans affects the default rate
5. To determine if the marital status of the applicant affects the default rate
6. To determine if the amount borrowed affects the applicant's ability to repay
7. To determine if the income of the applicant affects the default rate
8. To determine the relationship between the contract type of the loan and the loan default

Data Mining Success Criteria

We did explanatory analysis in order to understand how the various factors affected the client ability to repay the loans

2. Data Understanding

Data Understanding Overview

For this project, we are using the dataset from the kaggle website. These datasets are;

- Train Dataset - This dataset contains the different demographic information that are collected when a client is applying for a loan.
- Data Dictionary - This dataset contains the description of the Train dataset

Train Dataset - This dataset contains demographic information that is collected by the NBFIs . It has 121856 rows and 40 columns.

Data Dictionary- This dataset contains the description of the Train dataset and the column definitions.Some of the attributes are as follows:

- Client_Income- Client Income in $
- Car_Owned- Any Car owned by client before applying for the loan for another car (0 means No and 1 means otherwise)
- Bike_Owned - Any bike owned by client (0 means No and 1 means otherwise)
- Active_Loan- Any other active loan at the time of application of loan (0 means No and 1 means otherwise)

- House_Own- Any house owned by client (0 means No and 1 means otherwise)
- Child_Count- Number of children the client has
- Credit_Amount- Credit amount of the loan in $
- Loan_Annuity- Loan annuity in $
- Accompany_Client- Who accompanied the client when client applied for the loan
- Client_Income_Type- Clients income type
- Client_Education- Highest level of education achieved by client
- Client_Marital_Status-Marital status of client (D- Divorced, S- Single, M- Married, W- Widowed)
- Client_Gender- Gender of the Client
- Loan_Contract_Type - Loan Type (CL- Cash Loan, RL- Revolving Loan)
- Client_Housing_Type- Client Housing situation
- Age_Days- Age of the client at the time of application submission
- Employed_Days- Days before the application, the client started earning
- Registration_Days- Days before the loan application, the client changed his/her registration
- ID_Days- Days before the loan application, the client changed his/her identity document with which the loan w...
- Own_House_Age- Age of Client's house in years
- Client_Occupation-Client Occupation type
- Client_Family_Members- Number of family members does client have

3. Data Preparation

These are the steps followed in preparing our data:

1. **Loading data**

We imported pandas libraries to be used in loading our dataset from CSV files and created a data frame to be used for our analysis.

2. **Cleaning the data**

   We first removed spaces in the columns and converted them to lower case. For the missing values, we replaced them with mean for the numerical variables and mode for the categorical variables. We dropped the columns we did not need and the ones that contained a lot of missing values. We converted columns to appropriate data types and dropped the duplicates.

4. Analysis

- To determine if the gender of the applicant affects the default rate

    - Most of the people who applied for the loan were male, female applicants were few. Male applicants had the highest default rate while female applicants had a lower default rate compared to male applicants.

- To find out how the occupation of the applicant affects the default rate
    - Majority of the people who defaulted the loan did not indicate their occupation followed by laborers

- To determine if the age of the applicant affects the default rate
    - Most of the defaulters were are between the ages of 30 and 35

- To determine if having different loans affects the default rate
    - Clients who have a different loan are more likely to default compared to those who do not
- To determine if the marital status of the applicant affects the default rate
    - Clients who are married are more likely to take and to default on the loan

- To determine if the amount borrowed affects the applicant's ability to repay
    - Clients who borrow between 25000 and 50000 are the most likely to default

- To determine if the income of the applicant affects the default rate
    - Clients with lower income (between 0 and 20000) are most likely to default

- To determine the relationship between the contract type of the loan and the loan default
    - Clients taking cash loans are more likely to default

5. Data Modeling

To make predictions in our data we used tree-based models.
Below are the models that we used;

- Decision tree,
- Random forest,
- AdaBoost,
- Gradient boost models

We decided to use tree-based models because our dataset is non-parametric and our classes are binary. The models will also help us determine the most important features.

From the decision tree model findings the accuracy was 85% and the F1 score for the not defaulted was 94% and for the defaulted was 19%.

Accuracy from the random forest was 92% and the F1 score was 96% for the not defaulted and 12% for the defaulted.
From the AdaBoost and gradient boost models the accuracy was 91% and 92% respectively
The F1 score for our models was quite low though the accuracy was quite high and this was due to the class imbalance between those who defaulted and those who never defaulted hence we balanced our data in order to improve the performance

The performance after the balancing was:

- Decision Forest
  Accuracy = 92%
  F1 score
  Defaulted = 84%
  Not defaulted = 94%
- Random Forest
  Accuracy = 98%
  F1 score
  Defaulted = 97%
  Not defaulted = 99%
- AdaBoost
  Accuracy = 78%
- Gradientboost
  Accuracy = 78%

After the balancing our model performance improve except for the gradient boost and Adaboost whose performance dropped

5. Conclusion and Recommendations

Recommendations

- Target customers with rolling loans because they have a lower default rate
- Target customers with income above $ 20,000 because they have a lower default rate
- Target customers who borrow amount greater than $ 50,000 because they have a lower default rate
- Avoid lending loans to clients who do not indicate their occupation because they have a higher default rate
- Target customers who do not have an active loan
- Perform background checks for clients of ages between 25 and 45 because they have a higher default rate
- Check on external validation using score source variable column as it is the most important determinant of the default rate.

- Perform background checks for clients who are married because they have a higher default rate

Conclusion

The best model for predicting whether the client will default or not is the Random Forest with an accuracy of 98% and the best overall f1 score of .98 which was better than the other models used to make predictions.