**TITLE:** “Credit Card Default Prediction in R”

**AUTHORS:** Vaibhavi Panchal,Millicent N.O Boadu, and Cinoo Bosco Thomas.

**ACKNOWLEDGMENT:** This dataset was originally published alongside the 5th edition of William Greene’s book Econometric Analysis.

**INTRODUCTION:** The ability to accurately predict credit card defaults is crucial for financial institutions to manage credit risk effectively. In this project, we explore the use of machine learning techniques to develop predictive models for identifying credit card holders who are likely to default on their payments. We utilize the Credit Card dataset, which contains various attributes of credit card holders, including demographic information and credit history.

**DESCRIPTION OF DATA:** The Credit Card dataset from the AER package consists of information related to credit card holders. It includes several features that can be utilized for predictive modeling. In total, the data frame contains 1,319 observations on 12 variables. Below is a brief description of the features included in the dataset:

* **card:** [Categorical - Binary] Factor. Was the application for a credit card accepted?
* **reports:** [Numerical] Number of major derogatory reports.
* **age:** [Categorical/ Numerical] Age in years plus twelfths of a year.
* **income:** [Numerical] Yearly income (in USD 10,000).
* **share:** [Numerical] Ratio of monthly credit card expenditure to yearly income.
* **expenditure:** [Numerical] Average monthly credit card expenditure.
* **owner:** [Categorical - Binary] Factor. Does the individual own their home?
* **selfemp:** [Categorical - Binary] Factor. Is the individual self-employed?
* **dependents:** [Categorical] Number of dependents.
* **months:** [Numerical] Months living at current address.
* **majorcards:** [Numerical] Number of major credit cards held.
* **active:** [Numerical] Number of active credit accounts.

**NEEDED PACKAGES :**

library (AER)

library(ggplot2)

library(caTools)

library(rpart)

library(rpart.plot)

library(randomForest)

library(caret)

data("CreditCard")

**EXPLORATORY DATA ANALYSIS**

**1. Distribution of Credit Card Status**

ggplot(CreditCard, aes(x = card)) +

geom\_bar(fill = "blue", alpha = 0.7) +

labs(title = "Distribution of Credit Card Status",

x = "Card Status (0 = No Default, 1 = Default)",

y = "Count") +

theme\_minimal()

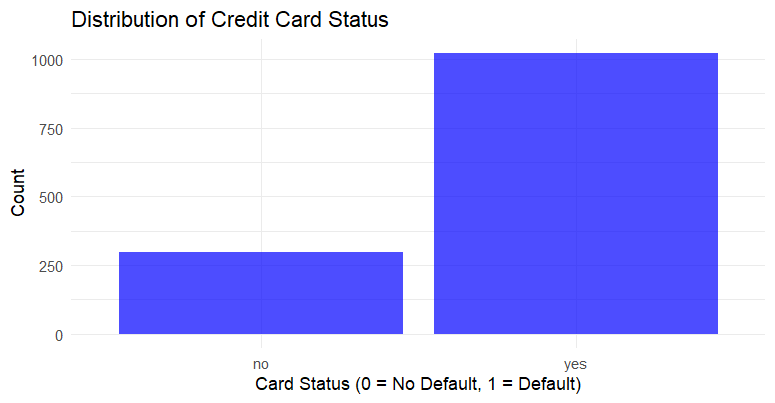


Fig 1 Distribution of Credit Card status

Here, we see the distribution of the target variable, credit card status. As you can see, most of the cardholders (represented by the taller bar on the left) did not default on their payments. This initial exploration helps us understand the overall balance of classes in our data.

**2. Distribution of Age**

ggplot(CreditCard, aes(x = age)) +

geom\_histogram(binwidth = 2, fill = "orange", color = "black", alpha = 0.7) +

labs(title = "Distribution of Age",

x = "Age",

y = "Count") +

theme\_minimal()

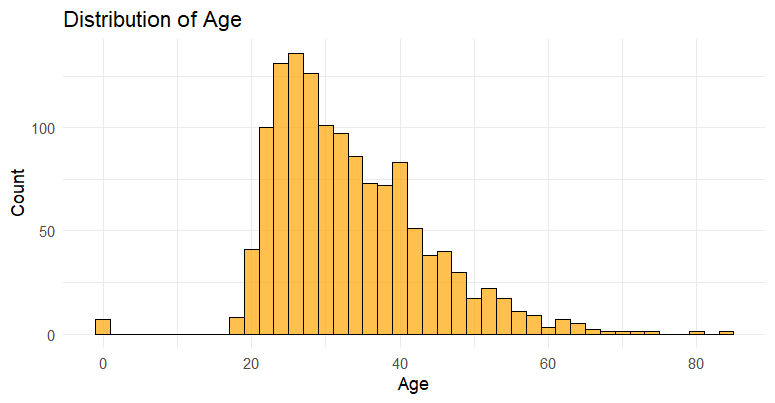


Fig 2 Distribution of Age

Fig 2, indicates that the age distribution within the dataset is varied, with the highest frequency of individuals falling within the age range of 27-29. The distribution appears to skew slightly towards older ages, with decreasing frequencies as age increases.

**3. Distribution of Income**

ggplot(CreditCard, aes(x = income)) +

geom\_histogram(binwidth = 1, fill = "green", color = "black", alpha = 0.7) +

labs(title = "Distribution of Income",

x = "Income",

y = "Count") +

theme\_minimal()

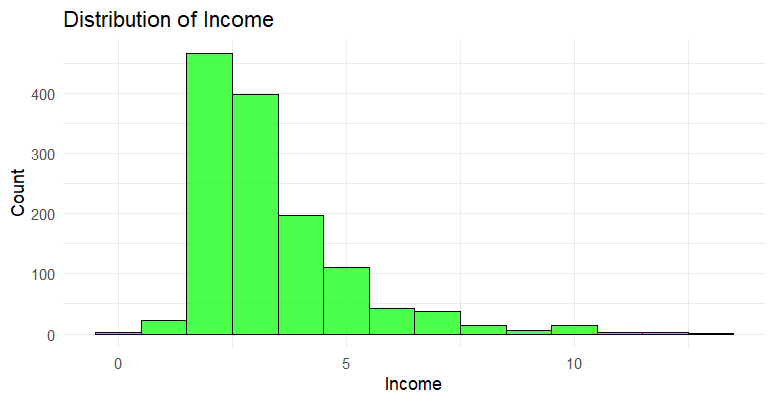


Fig 3 Distribution of Income

Fig 3, indicates that the income distribution within the dataset **creditcard\_clean** is varied, with the highest frequency of individuals having an income between $1000 and $1999. The distribution appears to skew slightly towards lower income brackets, with decreasing frequencies as income increases.

**4. Distribution of Reports**

ggplot(creditcard\_clean, aes(x = reports)) +

geom\_histogram(binwidth = 1, fill = "purple", color = "black", alpha = 0.7) +

labs(title = "Distribution of Reports",

x = "Number of Reports",

y = "Count") +

theme\_minimal()

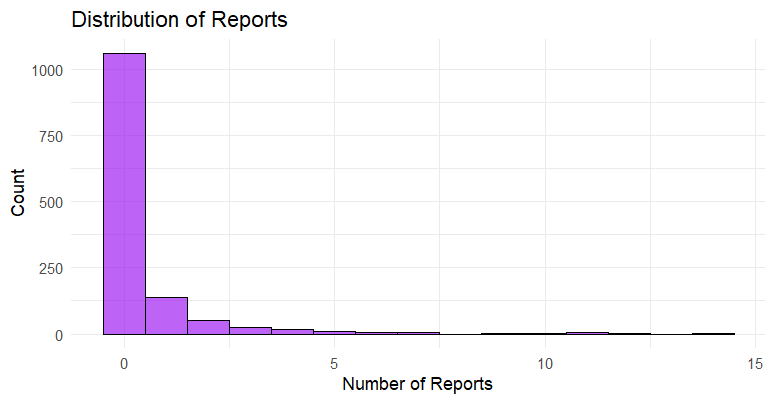


Fig 2.4 Distribution of Reports

Fig 2.4, illustrates that the majority of individuals in the dataset have fewer reports, with the highest frequency of individuals having 0 to 4 reports. As the number of reports increases, the frequency of individuals decreases, indicating a decreasing trend in the number of reports filed.

**METHOD(S)**

We employed two machine learning algorithms: decision trees and random forests, to build predictive models for credit card default prediction. Grid search is used to find the optimal complexity parameter (cp).

1. **Decision Tree Model:**

The decision tree is constructed by repeatedly partitioning the data based on predictor variables, creating decision nodes at each partition. The decision tree evaluates each predictor variable to determine the best split at each node. For example, it may ask questions such as "Does the person have more than 2 reports?" to decide how to partition the data effectively.

**A diagram of reports

Description automatically generated**

* **Model Performance:**

Calculate and print confusion matrix for decision tree model on testing data

|  |  |  |
| --- | --- | --- |
| Prediction Tree | NO | YES |
| NO | 32 | 5 |
| YES | 57 | 302 |

* **Decision Tree Model Evaluation Metrics:**

|  |  |
| --- | --- |
| Accuracy | 84.34 % |
| Precision | 35.96 % |
| Recall | 86.49% |
| F1 Score | 50.79 % |

**2. Random Forest Model:**

Random Forest, is used to build a predictive model for credit card ownership. It optimizes the model's performance by tuning hyperparameters through grid search and trains the model using the best hyperparameters obtained.

* **Random Forest Model Confusion Matrix (Testing Data)**

|  |  |  |
| --- | --- | --- |
| Prediction Tree | NO | YES |
| NO | 40 | 19 |
| YES | 49 | 288 |

* **Evaluation metrics for the random forest model.**

|  |  |
| --- | --- |
| Accuracy | 82.83 % |
| Precision | 44.94 % |
| Recall | 67.8 % |
| F1 Score | 54.05 % |

* **Feature Importance:**

A graph with numbers and a line

Description automatically generated with medium confidence

This plot shows the importance of each predictor variable in the Random Forest model. It indicates how much each variable contributes to reducing impurity in the decision trees of the ensemble. Higher values represent more important predictor variables, while lower values indicate less importance.

* **RESULT**

When comparing both the models the Decision Tree model achieves a higher accuracy(84.34%) and recall (86.49%)compared to precision(35.96%). This suggests that the model is good at identifying positive cases (defaults), but it also tends to predict a higher number of false positives. Whereas, the Random Forest model has slightly lower accuracy(82.83%) compared to the Decision Tree model. However, it has a higher precision,(44.94%) which means it is better at correctly identifying true positive cases with fewer false positives. The recall(67.8%) is lower than the Decision Tree model, indicating it may miss some actual positive cases. Nonetheless, the F1-Score(54.05%) is higher than the Decision Tree model, which shows a better balance between precision and recall.

* **RECOMMENDATIONS**

**Minimizing False Positives:**

Choose the Random Forest model for its higher precision. It will help avoid misclassifying customers who will not default as defaulters.

**Minimizing False Negatives:**

Choose the Decision Tree model for its higher recall. This model will help identify more potential defaulters, thereby minimizing financial risk.

**Model Interpretability:**

Choose the Decision Tree model for its straightforward and easy-to-understand structure. This can aid in explaining the model’s predictions to stakeholders and customers.

* **CONCLUSION**

Overall, each model has its own advantages. Depending on your specific objectives, such as minimizing false positives, minimizing false negatives, or prioritizing model interpretability, either the Random Forest or Decision Tree model could be a better fit for your Credit Card dataset analysis.

**REFERENCES:**

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Springer Science & Business Media.

[2] Bierman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[3] Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1-26.