Financial Profiles: A Study of Mortgage Solvency

By Caitlin Nguyen

Introduction

- This project analyzes mortgage solvency using a dataset provided by Shielder Consulting Company. The primary objectives are to identify factors leading to mortgage defaults and predict the total income of clients.

Background

Since its inception in 2010, Shielder Consulting Company has distinguished itself in the finance sector through its specialized focus on mortgage default analysis. With a robust track record of providing actionable insights, Shielder has become the go-to firm for banks grappling with loan solvency issues.

In the wake of a troubling uptick in mortgage defaults, a prominent bank sought Shielder's expertise. This surge in defaults was eroding the bank’s profitability, compelling them to seek a methodical approach to identify and address the underlying factors contributing to the escalating default rates.

The Task

Loan Default Prediction:

- Import the dataset and identify the 'TARGET' variable (loan default).
- Split the data into training (80%) and testing (20%) sets.
- Train a decision tree model to predict loan defaults.
- Visualize the decision tree and evaluate its accuracy using a confusion matrix.
- Discuss the primary factors influencing loan defaults.

Income Prediction:

- Predict the total income ('AMT_INCOME_TOTAL') of clients.
- Implement lasso and ridge regression models.
- Train these models on the scaled training set.
- Calculate the Root Mean Squared Error (RMSE) on the validation set.
- Assess significant predictors and discuss potential implications.

Data Overview

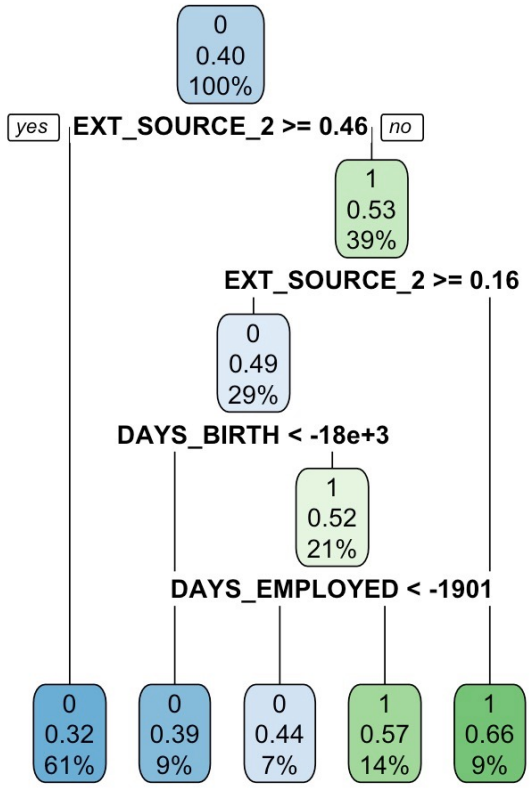
- The dataset includes over 34,000 active mortgages, containing client profiles, demographic details, and credit reports. Key variables include:

1. **TARGET**: Indicates whether a mortgage loan defaults (1) or not (0).
2. **AMT_INCOME_TOTAL**: The total income of the client.
3. **AMT_CREDIT**: The credit amount of the loan that the client has taken out.
4. **AMT_ANNUITY**: The loan annuity, a fixed sum of money paid to someone each year.
5. **NAME_CONTRACT_TYPE**: Indicates whether the loan is cash or revolving.
6. **CODE_GENDER**: The gender of the client.
7. **DAYS_BIRTH**: The age of the client in days at the time of the application.
8. **... and other relevant features listed in the appendix**.

Data Preprocessing

The dataset underwent several preprocessing steps:

1. **Handling missing values:** Rows with missing values were removed.
2. **Scaling:** Numerical features were centered and scaled.
3. **Partitioning:** The data was split into training and testing sets with an 80/20 split ratio.



Modeling: Decision Tree

- A decision tree model was trained to predict loan defaults. The data was partitioned into training (80%) and testing (20%) sets. The decision tree was visualized to understand the important features contributing to loan defaults.
- The decision tree model was effective in identifying key factors influencing mortgage defaults. The most significant predictor was the external credit score (EXT_SOURCE_2), followed by the client's age (DAYS_BIRTH) and employment duration (DAYS_EMPLOYED). The model achieved an overall accuracy of 63.94%, indicating a moderate ability to distinguish between default and non-default cases.

Model Evaluation: Decision Tree

- The decision tree model was evaluated using a confusion matrix. The model achieved an accuracy of 63.94%.

- **Confusion Matrix:**

  Reference

Prediction 0 1

  0 3467 1843

  1 640 935

- The confusion matrix revealed a higher number of false negatives, suggesting the need for improvements in capturing default risks.

Model Evaluation: Regression

- To predict the total income of clients (AMT_INCOME_TOTAL), both lasso and ridge regression models were employed. The models were trained on the scaled training set, and their performance was evaluated using Root Mean Squared Error (RMSE).
- The RMSE for the regression models were:

Lasso Regression: 76,361.94

Ridge Regression: 76,351.90

- The regression analysis for predicting client income utilized both lasso and ridge regression, with RMSE values of 76,361.94 and 76,351.90 respectively, demonstrating the models' potential in income prediction.
- Both models showed similar performance, indicating the need for further tuning or alternative modeling approaches.

Significant Predictors

- In the lasso regression model, certain coefficients were reduced to zero, highlighting the least significant factors. The significant predictors included features like the client's age, credit amount, and annuity.

Implications:

- These significant predictors can help banks to better assess the risk associated with mortgage applicants and make informed decisions.

Conclusion

- This analysis provided insights into the factors influencing mortgage defaults and client income prediction. The decision tree model highlighted key features driving defaults, while the regression models identified significant predictors of client income. Future work could involve exploring alternative models and feature engineering to improve prediction accuracy.