UNIVERSIDAD DISTRITAL FRANCISCO
JOSE DE CALDAS
FACULTAD DE INGENERIA

Machine Learning for Home Credit Risk Prediction

Fabian Esteban Perilla Lugo ♠ Julián Alejandro Pinzon Torres ♠ Daniel Alejandro Chavez Bustos ♠
♠20212020102 ♢20212020085
♡20212020109

Introduction

Granting mortgage loans is one of the most critical activities for financial institutions. However, the risk associated with these loans can be considerable, especially if the applicants are unable to meet their payment obligations. The consequences of nonpayment include significant financial losses for institutions and financial problems for borrowers.

To address this challenge, there is a need to develop a system that can effectively predict the probability of a mortgage loan applicant defaulting on their payments. This is where we will use data analysis and machine learning.

Goal

The main goal is to develop a machine learning model that can predict the likelihood of a loan applicant defaulting on their payments. With this tool, Home Credit can make more informed decisions about loan approvals, thereby reducing financial risk and improving the accuracy of evaluations for applicants with limited credit histories.

1 Experimental Setup

For this project, we used a Home Credit dataset containing information about previous loan applicants, their demographics, and loan performance. The experimental setup involved:

1.1 Data Preprocessing

- **Handling Missing Values:** Dropped columns with many missing values and replaced anomalies, particularly in 'CODE_GENDER'.
- **Encoding Categorical Variables:** Used label encoding for binary variables and one-hot encoding for others.
- **Standardization:** Standardized numerical features with `StandardScaler`.

1.2 Model Development

- **Neural Network:** Built with TensorFlow, using an input layer, multiple ReLU-activated dense layers, Dropout layers, and a sigmoid-activated output layer.
- **Compilation:** Compiled with Adam optimizer, binary cross-entropy loss, and accuracy metric.

1.3 Training and Validation

- Trained on processed data with 20% reserved for validation to monitor performance and prevent overfitting.
- Implemented early stopping when validation loss stopped improving.

1.4 Prediction

Used the trained model to predict default probabilities for the test dataset and created a submission file.

2 Experiments & Methodology

Our experiments focused on identifying effective preprocessing techniques and neural network architecture for predicting loan defaults. The methodology included:

2.1 Initial Data Analysis

Identified columns with significant missing values, high cardinality categorical variables, and potential outliers to inform preprocessing steps.

2.2 Data Cleaning and Encoding

- Dropped columns with '_AVG', '_MODE', and '_MEDI' suffixes due to missing values.
- Replaced anomalies in 'CODE_GENDER'.
- Label-encoded binary variables and one-hot encoded others.

2.3 Anomaly Removal

Used Z-score to remove outliers in 'DAYS_BIRTH', 'DAYS_EMPLOYED', and 'AMT_INCOME_TOTAL'.

2.4 Model Architecture Experiments

Tested various neural network architectures by adjusting layers, neurons, and Dropout layers. Settled on ReLU for hidden layers and sigmoid for the output layer.

2.5 Hyperparameter Tuning

Tuned hyperparameters (batch size, learning rate, epochs) with early stopping to prevent overfitting.

2.6 Evaluation Metrics

Evaluated performance with accuracy and ROC-AUC score on the validation set.

2.7 Final Model Training

Trained the best model configuration on the full training set (excluding the validation split) and generated predictions for the test dataset.

3 Conclusions

This project aimed to predict the likelihood of loan default by leveraging the dataset from the Kaggle competition "Home Credit Default Risk". Through a systematic approach encompassing data preprocessing, model development, and evaluation, several key insights and outcomes were achieved:

3.1 Effective Preprocessing Techniques

Proper handling of missing values and encoding of categorical variables proved essential in preparing the dataset for machine learning models. Standardizing numerical features ensured that all variables contributed equally to the model, improving overall performance.

3.2 Neural Network Performance

The neural network model, developed using TensorFlow, demonstrated the ability to capture complex patterns in the data. By experimenting with various architectures and incorporating techniques like dropout layers and early stopping, the model was fine-tuned to prevent overfitting and achieve robust predictions.

3.3 Model Evaluation and Metrics

The model's performance was evaluated using accuracy and ROC-AUC scores, providing a clear indication of its effectiveness in classifying loan defaults. The use of early stopping further ensured that the model did not overfit, maintaining its generalizability to new data.

3.4 Business Implications

Accurate prediction of loan defaults has significant implications for financial institutions. By identifying high-risk applicants, lenders can make informed decisions, potentially reducing the number of defaulted loans and enhancing profitability. This project demonstrated the feasibility of using machine learning models to support such decision-making processes.

3.5 Future Work

Future improvements could include exploring other advanced machine learning models, such as ensemble methods, and incorporating additional features that may further enhance predictive accuracy. Continuous evaluation and adaptation of the model in real-world scenarios will be crucial for maintaining its relevance and effectiveness.

Overall, this project highlighted the potential of machine learning in credit risk assessment, offering valuable insights and a solid foundation for further research and application in the financial sector.

- Pellentesque eget orci eros. Fusce ultricies, tellus et pellentesque fringilla, ante massa luctus libero, quis tristique purus urna nec nibh.
- Vestibulum sem ante, hendrerit a gravida ac, blandit quis magna.

References

[1] Kaggle. *Home Credit Default Risk Competition*. Available in: `https://www.kaggle.com/c/home-credit-default-risk`

[2] Home Credit Group. *About Home Credit Group*. Available in: `https://www.homecredit.net/`

[3] TensorFlow. *TensorFlow Documentation*. Available in: `https://www.tensorflow.org/`