MGMT635854-Data Mining&Anal for Mngrs

Project - I Report

The CRISP-DM data mining process – Statlog (German Credit Data)

Group Number: 05

Group Members:
Varun Rahul Mikkilineni(vm597)
Sai Mounika Koppolu (sk3368)
Pavithra Bobbala (pb543)
Arzu Senturk (as4629)

Business Understanding

Banks and credit card companies need to be careful when lending money to people. They must decide who is likely to pay back a credit card or loan and who is not. This is important because it helps companies avoid losing money when customers cannot pay back their debts. By using data mining, which means closely examining information to find patterns or trends, companies can makebetter decisions. This way, they can figure out who is a good person to lend money to.

Data mining helps find useful information in the data that companies have already collected. This can make it easier to understand customers' behaviors and financial habits. For example, by looking at a person's past spending and payment habits, a company can guess if the person will pay back in the future. Data mining helps banks and credit card companies to be more accurate intheir decisions and reduce the chances of lending money to someone who won't be able to pay it back. This keeps the business strong and allows it to continue lending to more people.

Data Understanding

The dataset provided comprises a total of 1000 entries and 21 columns, representing various attributes pertinent to consumer credit risk assessment. Each entry corresponds to an individualcustomer's information, encompassing a mix of both numerical and categorical variables.

- Numerical variables include attributes such as 'duration', 'amount', 'inst_rate', 'residing_since', 'age', 'num_credits', and 'dependents', which offer quantitativeinsights into the credit profiles.
- Categorical variables, including 'checkin_acc', 'credit_history', 'purpose', 'saving_acc', 'present_emp_since', 'personal_status', 'other_debtors', 'property', 'inst_plans', 'housing','job', 'telephone', and 'foreign_worker', provide qualitative context, elucidating various aspects like employment status, personal attributes, and credit history specifics.

The 'status' column, serving as the target variable, is binary, indicating the creditworthiness of the consumers — labeled as either 1 (Good) or 2 (Bad). This variable reflects the outcome thatthe predictive models will aim to forecast based on the information in the other columns.

Data Preparation

The preparation of data before fitting into a machine learning model is a critical step, ensuring that the dataset is structured, cleaned, and encoded to meet the model's input requirements. Our dataset underwent several stages of preprocessing aimed at tuning it for the most accuratepredictive outcomes.

Firstly, the target variable 'status', representing credit rating, was isolated from the feature set.Subsequently, dummy variables were created for the categorical attributes, expanding them into binary columns to make them suitable for the modeling process. For instance, the 'checkin_acc' column was split into multiple binary columns, each representing a specific category of the checking account, ensuring a numeric representation of categorical data.

Following this transformation, the response variable 'status' was converted into binary labels, where 'Good' is represented by 0 and 'Bad' by 1, thereby simplifying the prediction labels for the machine learning model. The dataset was then partitioned into a training set and a testingset, maintaining a 90-10 split. This division ensures that the model is trained on a diverse set of data and tested on unseen data to evaluate its performance accurately.

Scaling is another pivotal preprocessing step, where the data was standardized, meaning it was reshaped to a standard scale. This process involved using a StandardScaler to center and scale thedata, ensuring each feature contributes equally to the computation of distances in algorithms.

Additionally, a RandomForestClassifier was utilized to assess the importance of each feature,an approach that aids in understanding the contribution of each attribute toward the predictivemodel. Based on this importance score, features were selected or omitted, allowing for the enhancement of the model by focusing on the most impactful attributes. In our case, 35 features were selected for model fitting based on a predefined threshold of importance.

In conclusion, through various stages of data preparation, including encoding, scaling, andfeature selection, the dataset has been optimized, ensuring it is in an ideal format and structure for effective model training and evaluation.

Modeling

In this section, we will discuss the modeling phase where two data mining models, LogisticRegression and Random Forest Classifier, are utilized to solve our credit scoring problem.Various metrics such as accuracy, precision, recall, and F1-score are used to assess the model's performance. Additionally, hyperparameter tuning was conducted to improve the model's performance.

Two models were trained using the preprocessed and selected features:

- Logistic Regression Results:
  - The logistic regression model achieved an accuracy of 78% on the testing data.The precision, recall, and F1-score for predicting bad credits (1) were 0.62, indicating that there is a balanced performance in identifying the positive class.
- Random Forest Classifier Results:
  - The Random Forest model resulted in an accuracy of 72% on the test data.However, the precision and recall for identifying bad credits (1) were comparatively lower than the logistic regression model.

Hyperparameter Tuning:

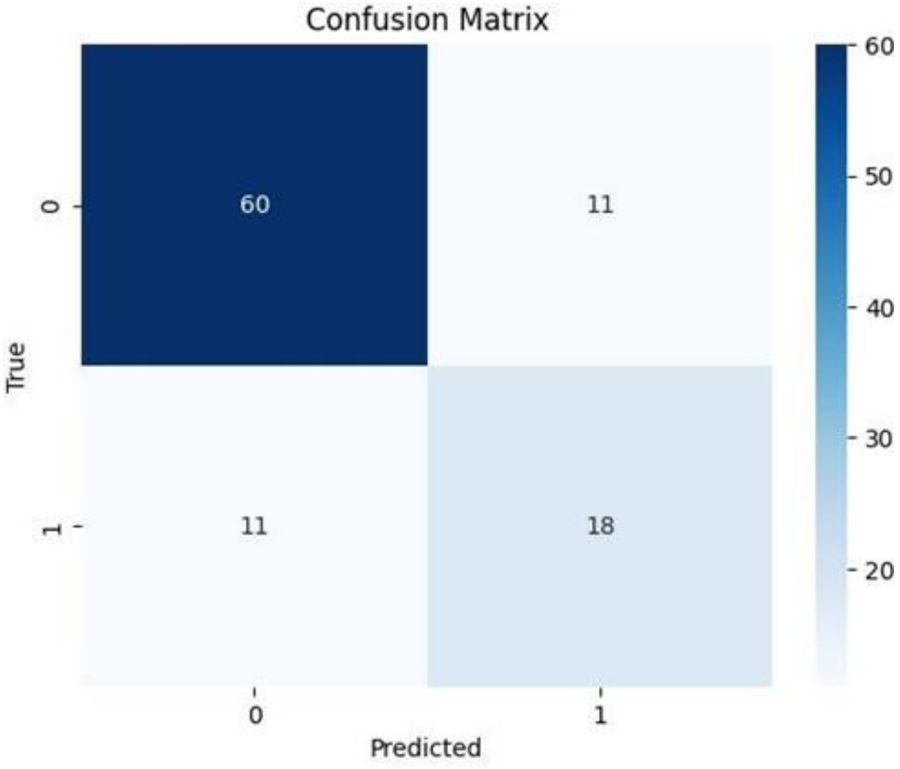
For the Logistic Regression model, hyperparameters such as the regularization constant (C), penalty, and solver were tuned using GridSearchCV, which performed an exhaustive search over the specified parameter values. The best parameters were {'C': 0.1, 'penalty': 'l2', 'solver':'liblinear'}, and applying these parameters didn't change the accuracy; it remained at 78%.

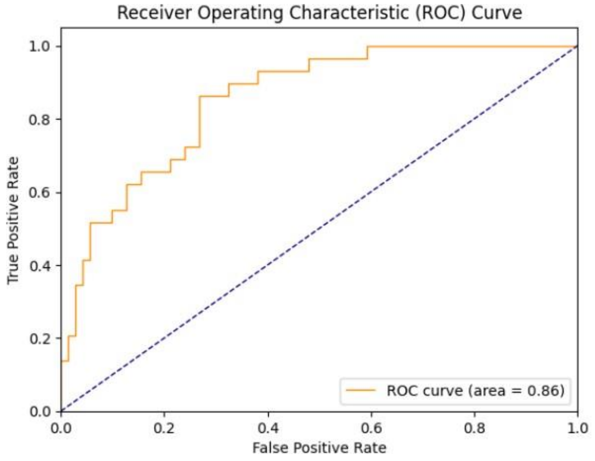
Visualizations:

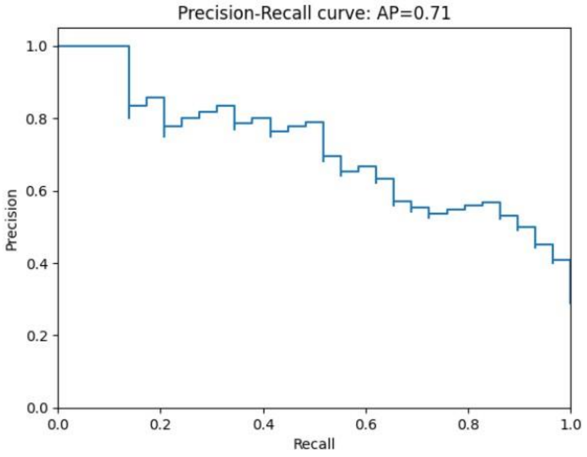
Confusion Matrix: The confusion matrix illustrated that the model correctly identified 60 good creditsand 18 bad credits, but there were 11 false positives and 11 false negatives. This insight is crucial as it shows where the model is making mistakes in terms of false positives and negatives.

ROC Curve: The ROC Curve demonstrated a stepwise increase, plateauing at an accuracyrate of 0.6. This indicates a reasonable balance between the true positive rate and false positive rate, but there's room for improvement.

Precision-Recall Curve: curve shows a trade-off between precision and recall for different thresholds. Curve is balanced, not leaning excessively towards precision or recall. This suggests that the model has a harmonized ability to correctly identify positive instances andlabel negative instances as negative, crucial for a credit scoring model where both false positives and false negatives carry significant costs.







<u>Conclusion on Modeling</u>:

The Logistic Regression model demonstrated a more balanced performance in identifying goodand bad credits compared to the Random Forest model. Hyperparameter tuning further refined the model, although the accuracy remained constant. Visualizations such as the confusion matrix and ROC curve provided essential insights into the model's performance, highlighting areas where improvements are necessary to make the model more reliable and robust for credit scoring.

Evaluations

The primary goal of our data mining endeavor was to accurately predict the creditworthiness of customers. The models, particularly the logistic regression, showed a promising 78% accuracy, but there's more to delve into beyond this. In terms of business success criteria:

- False Positives and False Negatives: The confusion matrix revealed that there wereboth false positives and negatives. In a credit scoring context, false negatives (bad loans classified as good) could lead to potential losses due to bad debt, while false positives could result in lost opportunities.
- Precision and Recall: Precision and recall, especially in identifying bad credits, were found to be balanced. This balance is significant in a business context, as both avoidingbad loans and identifying potential good loans are crucial.
- Robustness: The model should be robust enough to handle diverse customer profiles andchanging economic times. It might need further evaluation and tuning to ascertain this.

Considering these aspects, the data mining results show promise but also room for improvement. Deciding whether to implement these models directly would require a more comprehensive evaluation, including considerations like cost-benefit analysis and potentialimpacts on customer relationships.

Deployment

Benefits:

- Automated Decision-Making: Implementing the model could automate the creditscoring process, making it more efficient.
- Risk Mitigation: The model can help in identifying potential bad loans, thus reducingthe risk of bad debt.

Concerns and Issues:

- Model Interpretability: Logistic Regression is interpretable, but ensuring thatstakeholders understand and trust the model's decisions is essential.
- Data Privacy and Ethics: Ensuring that customer data is handled securely and ethicallyis paramount.
- Changing Economic Environments: The model should be resilient and adaptableto varying economic climates, which needs continuous monitoring and tuning.

Wisdom from the Project:

- Importance of Data Preparation: Significant insights were gained regarding the pivotalrole of data preprocessing and feature selection in improving model performance.
- Model Evaluation: A deeper appreciation was cultivated for holistic model evaluationbeyond accuracy, considering business impacts like false positives and negatives.
- Continuous Improvement: Recognizing the necessity for continuous model monitoring,evaluation, and improvement for adapting to evolving conditions and improving decision-making.