CREDIT CARD FRAUDULENT DETECTION

Recommendations and Key Outcomes

MAY 27, 2024

VINAY PAVAN GUNTUR

Contents

Recommendations and Key Outcomes

Objective

In this case study, aimed to develop a predictive model to accurately identify fraudulent credit card transactions using a given dataset. The key steps and findings of the task can be summarized as follows:

GOAL

The goal of this project was to build a robust machine learning model to accurately detect fraudulent transactions. After evaluating multiple models, the Random Forest classifier was chosen for its superior performance, particularly when combined with SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

Data Preprocessing:

- Handled missing values and encoded categorical variables appropriately.
- Split the dataset into features (independent variables) and the target variable (Fraud).

Exploratory Data Analysis (EDA):

- Analysed the distribution of the target variable and identified a significant class imbalance, with a very small proportion of transactions labeled as fraudulent.
- Explored the distribution of features and their relationships with the target variable.

Feature Engineering:

- Created new relevant features and normalized numerical features to improve model performance.

Model Selection and Evaluation:

- Trained and evaluated multiple Machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks.
- Addressed the class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

Key Findings:

- Models trained on data without addressing the class imbalance performed poorly in terms of F1 score, due to the overwhelming number of benign transactions.
- Applying SMOTE significantly improved the performance of all. The F1 scores increased notably, indicating better detection of fraudulent transactions.
- Among all models, the Random Forest Classifier with SMOTE achieved the highest F1 score **(0.983621)**, making it the most effective model for this task.

Optimization:

- The selected Random Forest model was further optimized using Grid Search Cross-Validation (GridSearchCV). This process involved fine-tuning hyperparameters to maximize the model's performance.

Key Hyperparameters Tuned:

- n_estimators: Number of trees in the forest.
- max_depth: Maximum depth of the tree.
- min_samples_split: Minimum number of samples required to split an internal node.
- min_samples_leaf: Minimum number of samples required to be at a leaf node.

Grid Search Results:

- The best combination of hyperparameters was identified through Grid Search, and the optimized model achieved an impressive **ROC AUC Score of 0.99**.

Model Evaluation:

- The ROC AUC score is a crucial metric for evaluating the performance of a classification model, especially in imbalanced datasets like fraud detection.
- It represents the area under the Receiver Operating Characteristic (ROC) curve and provides an aggregate measure of performance across all classification thresholds.

Interpreting the ROC AUC Score:

- The model's performance is equivalent to random guessing = 0.5
- The model perfectly distinguishes between fraudulent and non-fraudulent transactions = 1.0
- An ROC AUC score of **0.99** indicates excellent discriminatory power, meaning the model is highly effective at distinguishing between fraudulent and non-fraudulent transactions.

Evaluation Metrics:

In addition to the ROC AUC score, the optimized model's performance was further assessed using various metrics:

- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positive predictions among all actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a single metric to evaluate the balance between them.
- Confusion Matrix: A summary of prediction results, showing the number of true positives, true negatives, false positives, and false negatives.

Key Findings and Business Implications:

- High ROC AUC Score: The model's high ROC AUC score reflects its ability to effectively distinguish between fraudulent and legitimate transactions.
- Precision-Recall Balance: The balance between precision and recall ensures that the model minimizes false positives (non-fraudulent transactions flagged as fraudulent) while maximizing the detection of actual fraudulent transactions.
- Confusion Matrix Insights: The confusion matrix provides insights into the model's accuracy and the types of errors it makes, aiding in further refinements and understanding of model performance.

Business Benefits:

- Improved Fraud Detection: By implementing this model, the organization can significantly enhance its fraud detection capabilities.
- Reduced False Alarms: Higher precision leads to fewer false alarms, improving customer experience.
- Minimized Financial Losses: Higher recall ensures that most fraudulent transactions are detected, reducing potential financial losses.

Recommendations:

- Implement the Random Forest Classifier with SMOTE in a real-time fraud detection system to enhance the accuracy of identifying fraudulent transactions.
- Continuously monitor and retrain the model to adapt to new patterns of fraudulent activity and maintain high performance over time.

Conclusion

- This task successfully demonstrated the importance of addressing class imbalance in fraud detection tasks.
- By applying SMOTE, and significantly improved the models' ability to detect fraudulent transactions.
- The Random Forest Classifier with SMOTE emerged as the best-performing model, offering a robust solution for real-time fraud detection.
- This approach not only enhances the accuracy but also provides a scalable method to keep up with evolving fraud patterns, ensuring the reliability and security of financial transactions.

Additional Steps for comprehensive coverage

Feature Importance:

- Analyze the importance of each feature in the Random Forest model. This can provide insights into which features contribute most to the model's predictions.

Cross-Validation:

- Ensure that cross-validation was performed correctly. This can help validate the model's performance across different subsets of the data.

Model Explainability:

- Use tools like SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) to explain the model's predictions. This can be crucial for gaining trust from stakeholders.