Predicting Credit Card Fraud

Amanda Altamirano, Liam Riener, Abhik Shrestha, Charlene Vu

DATA 300: Statistical and Machine Learning

INTRODUCTION

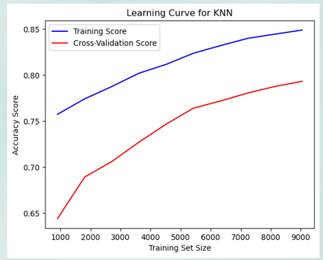
The research focuses on predicting credit card fraud using machine learning models, addressing a critical issue in financial services where fraudulent transactions pose significant risks to both institutions and customers. As credit card usage becomes increasingly prevalent, detecting and preventing fraud is essential to ensure financial security and maintain trust. This study employs various machine learning models to analyze patterns and behaviors indicative of fraudulent activities, enabling more efficient and accurate fraud detection systems.

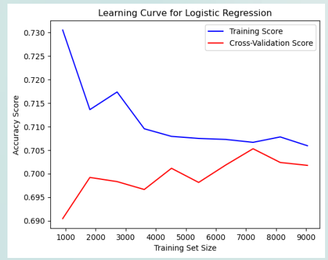
OBJECTIVES

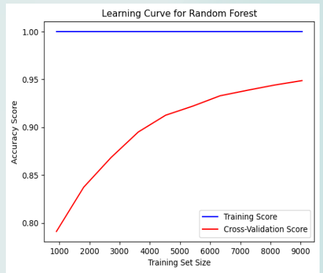
- To identify key features and patterns that differentiate fraudulent transactions from legitimate ones.
- To evaluate and compare the performance of various machine learning models, including K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes, and Random Forest, in predicting credit card fraud.
- To enhance the predictive accuracy of fraud detection systems by addressing class imbalance and optimizing model parameters.

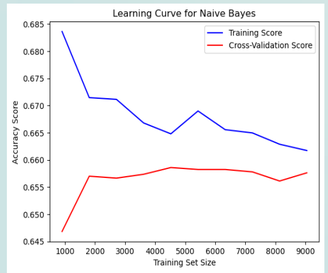
DATA AND METHODOLOGY

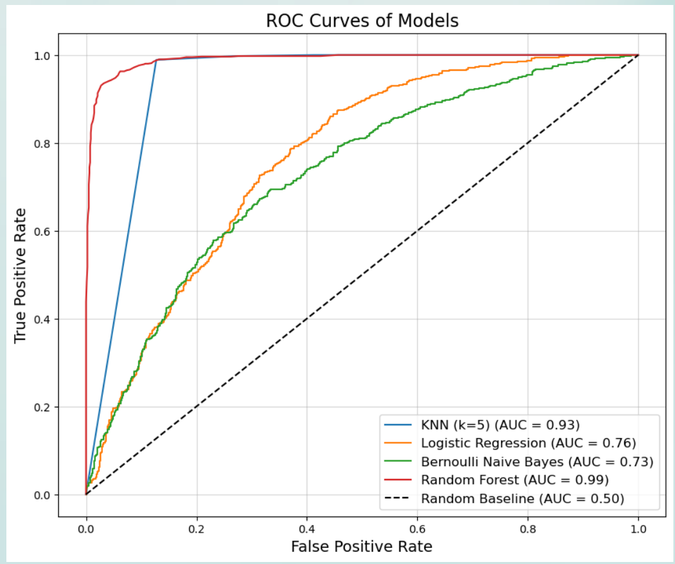
- Data overview: The dataset, sourced from Kaggle, initially contained 307,510 rows and 122 features. It included missing values in 67 variables and 298,909 rows, presenting significant challenges for data preprocessing. After thorough cleaning and feature selection, the final dataset was trimmed to 8,602 observations and 48 features.
- Scaled all features to standardize their ranges, ensuring equal prioritization during modeling.
- The dataset exhibited a significant imbalance, with fraudulent transactions comprising only 1% of the data initially. After removing rows with missing values, this percentage increased to 6%. SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset, achieving a 50:50 ratio of fraudulent to non-fraudulent transactions. There are 16,152 observations in the resampled set.
- Performed PCA on features: 48 features explaining 90% of variance.
- Split data into 70% training, 20% validation, 10% testing.
- Models: K-Nearest Neighbors (1,2,3,5,7,9,11 neighbors), Logistic Regression, Bernoulli Naïve Bayes, Random Forest

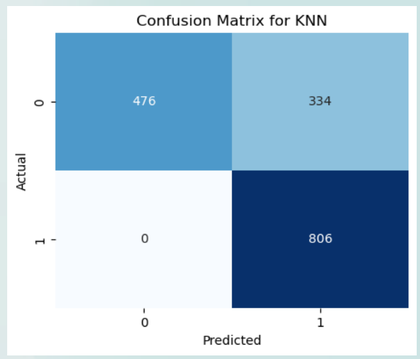


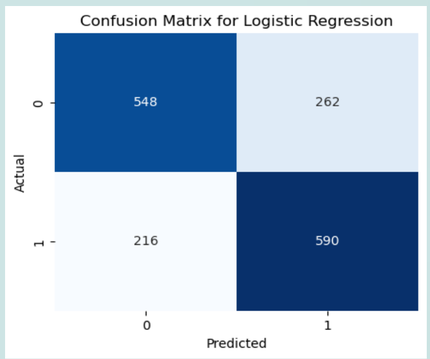


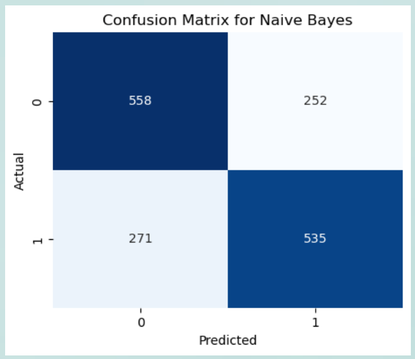


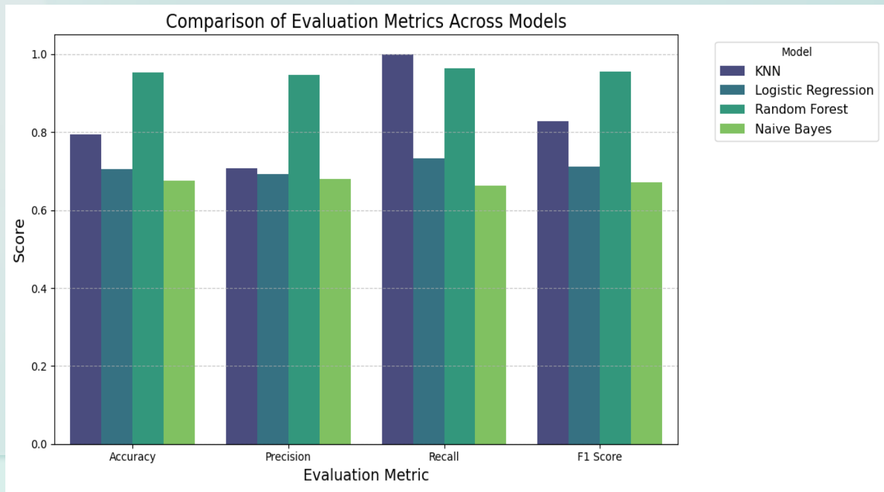


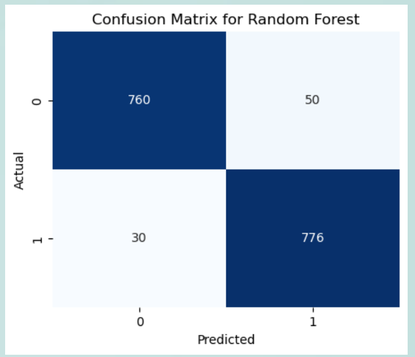












DISCUSSION

- **K-Nearest Neighbors (KNN)**: The KNN model achieved its best performance with k=2, yielding validation and testing accuracies of 91.15% and 90.72%. It achieved perfect recall for fraud cases, avoiding false negatives, but suffered from higher false positives, making it ideal for applications prioritizing fraud detection over minimizing false alarms.
- **Logistic Regression**: Logistic Regression, with accuracies of around 70%, served as a solid baseline but struggled to capture non-linear patterns and handle class imbalance and large dataset, resulting in significant false negatives. This limits its applicability for real-world fraud detection.
- **Naïve Bayes (Bernoulli)**: Naïve Bayes showed moderate accuracy (65% to 67%) and balanced precision and recall but often misclassified fraud cases due to its simplicity and sensitivity to class imbalance.
- **Random Forest Classifier**: Random Forest was the most accurate model, with validation and testing accuracies of 95.82% and 95.30%. Its balanced precision and recall across classes and ability to handle complex patterns make it the most reliable option, though its computational complexity may pose challenges for larger datasets.

CONCLUSION

- The best model is Random Forest Classifier with over 95% accuracy and balanced performance.
- KNN model is a viable alternative with perfect fraud recall but prone to false positives.
- Future work could focus on improving sampling and scalability, exploring advanced techniques, such as ensemble learning or deep learning, to further enhance predictive performance.

REFERENCE

- Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia Computer Science, 132,* 385–395. Elsevier.
- Mienye, D., Ibomoiye, & Jere, N. (2024). Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions. *IEEE Journals & Magazine*, IEEE Xplore, July 11, 2024.