**CAP 4770: Data Mining**

**Final Project Documentation**

**Corbin Lane**

**Edwin J Soto Lopez**

**Fall 2024**

Email: [cjl50@students.uwf.edu](mailto:cjl50@students.uwf.edu), [ejs47@students.uwf.edu](mailto:ejs47@students.uwf.edu)

**Abstract**

In this project, we try to address the problem of fraud detection in financial transactions using machine learning techniques. The dataset used for this study is a Credit Card Fraud Detection dataset, which contains anonymized credit card transactions made by European cardholders in September 2013. The dataset is highly imbalanced, with only 0.17% of fraudulent transactions. Our approach mainly uses two supervised learning algorithms. The two algorithms used are logistic regression and random forests to detect fraudulent transactions. We will preprocess the data by scaling the features, handling outliers, and applying SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. Various models will be compared based on performance metrics like precision, recall, and the area under the ROC curve (AUC), focusing on optimizing precision to reduce false positives. The goal is to develop a reliable system that can accurately identify fraudulent transactions while minimizing false alarms.

**Background**

Credit card fraud has become a pervasive issue in the digital age, posing significant financial risks to both individuals and institutions. As technology advances and digital transactions proliferate, fraudsters continually evolve their tactics to exploit vulnerabilities in payment systems. This research aims to address this pressing concern by leveraging machine learning techniques to develop a robust model capable of accurately detecting fraudulent credit card transactions.

The rapid growth of e-commerce and online payments has created a fertile ground for cybercriminals. Fraudulent activities, such as unauthorized purchases, identity theft, and account takeover, can lead to substantial financial losses for both consumers and businesses. To combat these threats, financial institutions and payment processors have implemented various security measures, including encryption, tokenization, and biometric authentication. However, these measures alone are often insufficient to prevent sophisticated fraud attacks.

Machine learning offers a powerful tool for detecting anomalous patterns in large datasets, making it well-suited for fraud detection. By analyzing historical transaction data, machine learning models can identify subtle indicators of fraudulent behavior that may be difficult to detect through traditional rule-based systems. Various machine learning algorithms, including decision trees, random forests, neural networks, and support vector machines, have been applied to credit card fraud detection with varying degrees of success.

One of the key challenges in credit card fraud detection is the inherent imbalance in the dataset. Fraudulent transactions typically represent a small fraction of the total transactions, leading to a class imbalance problem. This imbalance can negatively impact the performance of machine learning models, as they may be biased towards the majority class (legitimate transactions). To address this issue, techniques such as oversampling, undersampling, and synthetic minority oversampling technique (SMOTE) can be employed to balance the dataset.

Another important consideration in fraud detection is the evolving nature of fraudulent activities. Fraudsters constantly adapt their techniques to circumvent existing detection methods. Therefore, it is essential to continuously monitor and update the machine learning model to ensure its effectiveness. This can be achieved through techniques such as retraining the model with new data, incorporating feedback from human analysts, and adapting to changes in transaction patterns.

In conclusion, the application of machine learning to credit card fraud detection has the potential to significantly improve the accuracy and efficiency of fraud prevention systems. By leveraging advanced algorithms and techniques, it is possible to build robust models that can identify and mitigate fraudulent activities, safeguarding both individuals and institutions from financial loss.

***Section 1: Overview of the Problem***

Credit card fraud poses a significant financial threat to both individuals and institutions. The rapid growth of digital transactions has made detecting and preventing fraudulent activities increasingly challenging. This project aims to address this issue by developing a robust machine-learning model to identify fraudulent credit card transactions accurately. The specific problem we aim to solve is classifying credit card transactions as either fraudulent or legitimate. Given a dataset of historical transactions, we will train a model to predict the likelihood of a new transaction being fraudulent.

**Project Scope/Limitations**

Let's begin by addressing our project's scope and Limitations. The following is a list based on the dataset's specific characteristics, the chosen modeling approach, and the project’s focus on offline analysis rather than real-time implementation. The project scope is constrained by the dataset's age, anonymization, and class imbalance, as well as by the selection of supervised learning algorithms and the exclusion of live deployment and monitoring.

* **Dataset:** This project's scope is limited to the "creditcard.csv" dataset, which contains anonymized credit card transactions from September 2013.
* **Model Selection:** We will mainly use supervised learning algorithms, specifically logistic regression and random forests.
* **Imbalanced Dataset:** The dataset is highly imbalanced, with a significant disparity between fraudulent and legitimate transactions. Addressing this imbalance will be an essential aspect of the project.
* **Real-time Detection:** This project focuses on offline analysis and model development. Real-time deployment and continuous monitoring of fraud detection systems are beyond the scope of this initial investigation.

***Section 2: Summary of Approach***

To address the problem of credit card fraud detection, we will employ a machine learning approach. Our methodology involves the following steps:

**Data Preprocessing**

* **Exploratory Data Analysis (EDA):** Conduct a thorough analysis of the dataset to understand its characteristics, including data types, missing values, and distribution of features.
* **Data Cleaning:** Handle missing values, outliers, and inconsistencies in the data.
* **Handling Imbalanced Data:** Apply techniques like oversampling (SMOTE) or undersampling to balance the dataset and prevent model bias.

**Model Selection and Training**

* **Logistic Regression:** A simple yet effective model for binary classification.
* **Random Forest:** An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
* **Model Training:** Train each model on the preprocessed dataset, tuning hyperparameters to optimize performance.

**Model Evaluation**

* **Performance Metrics:** Evaluate the performance of each model using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).
* **Model Selection:** Choose the model with the best performance based on the selected metrics, considering the trade-off between precision and recall.

**Model Deployment:** For deploying the fraud detection model, we need to select an appropriate strategy that fits the operational needs. One option is to use a cloud-based solution, which supports handling large volumes of transactions and provides remote access to the model. Another option is to integrate the model directly into an existing fraud detection system, allowing it to work within the current infrastructure. The goal is for the model to process transactions in real-time to identify potential fraud cases.

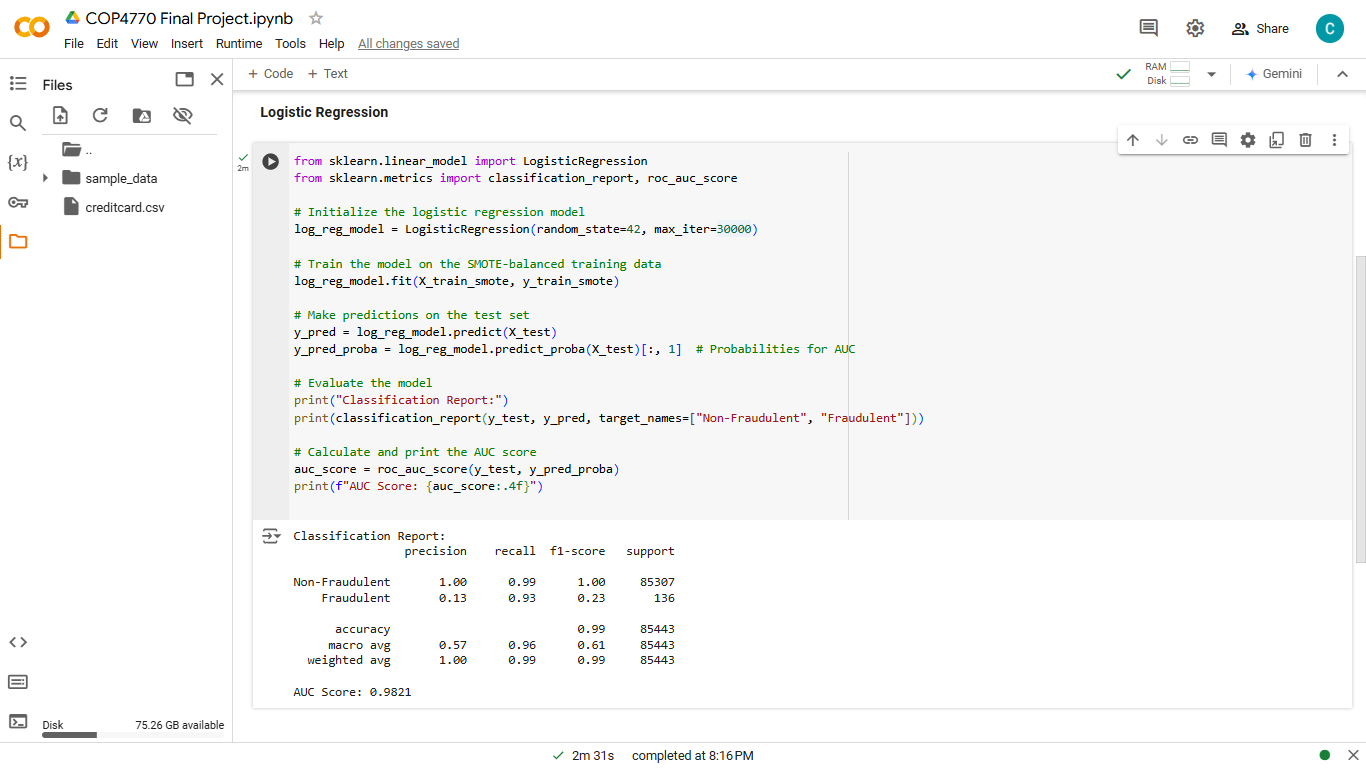
**Monitoring and Maintenance:** After deployment, monitoring the model’s performance is crucial to ensure accuracy over time. Since fraud patterns change, the model’s performance may decrease if it isn’t updated. Setting up a system to track its effectiveness and periodically retrain the model as new data becomes available helps keep it accurate and responsive to new types of fraud. This approach helps maintain the model’s reliability and effectiveness in detecting fraudulent transactions.

***Section 3: Demonstration***

Before we document the model’s capabilities let's review our *evaluation metrics*.The evaluation metrics help with assessing how well the model performs in identifying and distinguishing between classes. The following are evaluation metrics used for our models.

1. **Accuracy:** The percentage of total transactions correctly classified as either fraudulent or non-fraudulent by the model.
2. **Precision:** The percentage of transactions predicted as fraudulent that are actually fraudulent, reflecting the model’s accuracy in identifying frauds without false alarms.
3. **Recall:** The percentage of actual fraudulent transactions that the model successfully identifies, showing how well it captures all frauds.
4. **F1-score:** The harmonic mean of precision and recall, balancing both to provide a single performance measure, especially useful for imbalanced datasets.
5. **AUC (Area Under the Curve):** The measure of the model’s ability to distinguish between fraudulent and non-fraudulent transactions across various thresholds; higher values indicate better discrimination.

**The Logistic Regression Model**

****

**Model Initialization:** We initialize LogisticRegression with max\_iter=30,000 to allow more iterations than needed.

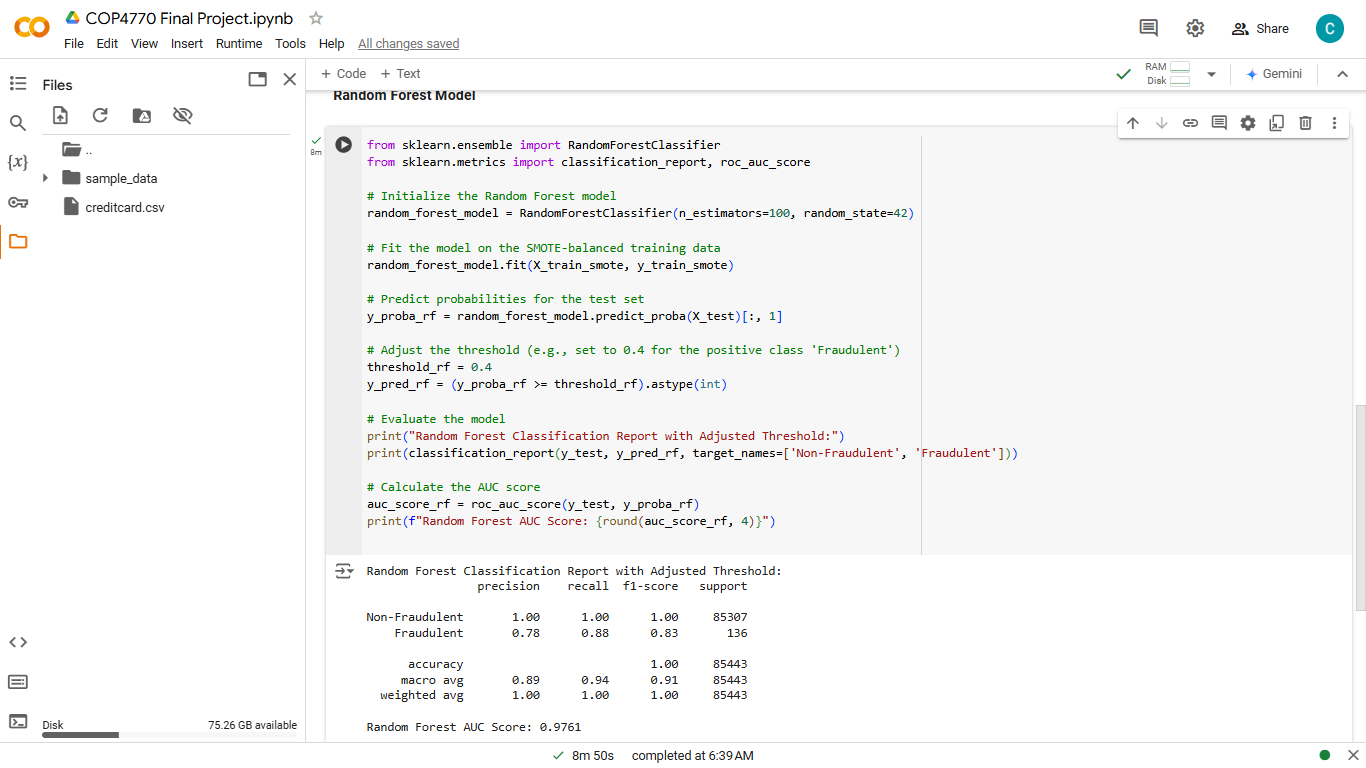
**Evaluation:** classification\_report provides metrics like precision and recall, which are needed for fraud detection, and roc\_auc\_score provides the AUC metric.

**High recall for Fraudulent transactions (0.93):** This is excellent since recall is crucial in fraud detection to capture most fraudulent cases.

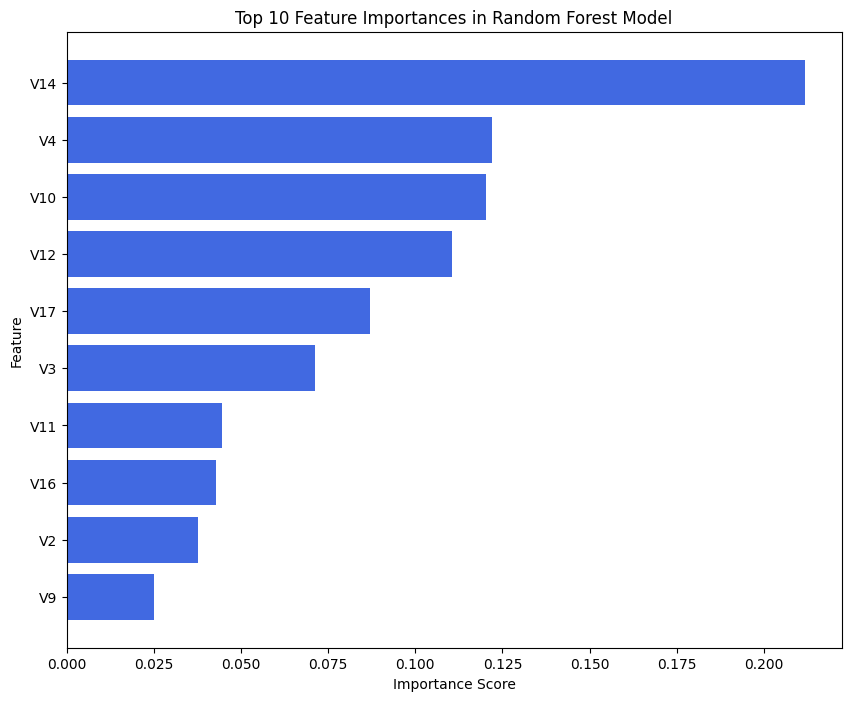
**Precision for Fraudulent transactions is still low (0.13):** This is expected in highly imbalanced datasets and indicates some false positives.

**Accuracy and AUC: Both remain high**: with accuracy at 0.99 and AUC at 0.9821, showing the model effectively differentiates between fraudulent and non-fraudulent cases.

**The Random Forest Classifier Model**



The Random Forest model has performed well, especially with the high recall and precision for the fraudulent class compared to the Logistic Regression model. Precision for the fraudulent class (0.78) and recall (0.88) indicate the model’s ability to accurately identify fraudulent transactions. AUC Score (0.9761) is close to the Logistic Regression score, but with better balance in precision and recall, suggesting improved detection capability with fewer false negatives.



This plot shows the top 10 features that the Random Forest model finds most valuable for distinguishing fraudulent from non-fraudulent transactions.



***Section 4: Conclusion***

The Logistic Regression model achieved an accuracy of 98.59%, indicating that it correctly classified the majority of transactions. Its recall for detecting fraudulent transactions is exceptionally high at 92.65%, meaning it successfully identifies almost all fraud cases. However, its precision for fraud is quite low at 9.55%, which means that a significant portion of flagged transactions are false positives (not actually fraudulent). The F1-score, which balances precision and recall, is also low at 0.1731, showing that while the model catches fraud effectively, it does so at the cost of misclassifying many legitimate transactions as fraudulent. The model’s AUC of 0.9824 suggests strong discriminatory power overall.

In contrast, the Random Forest model has a slightly higher accuracy of 99.94%, and its precision for fraudulent transactions is much better at 77.78%. This indicates a lower rate of false positives, making it more reliable when it flags a transaction as fraud. Its recall for fraud is also strong at 87.5%, meaning it captures a large portion of actual fraudulent cases, though slightly fewer than the Logistic Regression model. The F1-score of 0.8235 reflects a good balance between precision and recall for fraud detection. The AUC score of 0.9761 indicates that this model also performs well in distinguishing between fraudulent and non-fraudulent transactions. To back up this claim let's look at some visualizations. The following graphs enhance the understanding and comparison of the two model performances.

Despite how well our logistic regression model performed, the Random Forest model strikes a better balance between identifying fraudulent transactions and avoiding false positives, making it the stronger overall choice for fraud detection in this case.

**Key Lessons:**

1. **Data Quality and Preprocessing are Essential:** Preparing the data properly, especially when dealing with an imbalanced dataset, was critical to the project’s success. Techniques like SMOTE for balancing classes and feature scaling were essential to improve model performance and prevent biased predictions.
2. **Model Selection Depends on Trade-Offs:** Different models have strengths and weaknesses, and understanding the trade-offs between them—especially between precision and recall is crucial in fraud detection. We learned that while Random Forest offered high precision and recall, Logistic Regression provided insights into feature relationships, making each model useful depending on the specific requirements.
3. **Balancing Performance Metrics is Challenging:** Optimizing one metric often comes at the cost of another. For instance, improving precision can lower recall, which taught us the importance of selecting metrics that align with project goals in this case, reducing false positives while catching fraudulent transactions.
4. **Deployment Requires More than Just a Good Model:** Simply having a high-performing model isn’t enough. Deployment needs careful planning, including decisions around infrastructure, monitoring, and maintenance, to ensure the model remains effective over time.

**References**

KNOWLEDGE DOCTOR. (2024, April 25). *Project 22 : Credit Card Fraud Detection Using Machine Learning* [Video]. YouTube. https://www.youtube.com/watch?v=1oC1omuFLrI

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to data mining (Global edition.). Pearson.

Chowdhury, D., Hovda, S., & Lund, B. (2023). Analysis of cuttings concentration experimental data using exploratory data analysis. *Geoenergy Science and Engineering*, *221*, 111254-. https://doi.org/10.1016/j.petrol.2022.111254

Hong, H., Pourghasemi, H. R., & Pourtaghi, Z. S. (2016). Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology*, *259*, 105–118. https://doi.org/10.1016/j.geomorph.2016.02.012

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 4(1), 55–63. https://doi.org/10.1002/widm.1114

What is Logistic Regression? - Logistic Regression Model Explained - AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/logistic-regression/

Stojiljković, M. (2023, June 26). *Logistic Regression in Python*. https://realpython.com/logistic-regression-python/

*W3Schools.com*. (n.d.). https://www.w3schools.com/python/python\_ml\_logistic\_regression.asp