Credit Card Fraud Detection Report

HarvardX Data Science Professional Certificate: PH125.9x Capstone 2

Jennifer Cheng

November 4, 2024

Table of Contents

Credit Card Fraud Detection Report....1
1. Introduction....3
2. Methods/Analysis....4
2.1 Dataset and Variables....4
2.2 Data Wrangling....5
2.3 Feature Engineering....6
2.4 Data Exploration and Visualization....6
3. Modeling....12
3.1 Pre-Processing....12
3.2 Supervised Learning Models....12
3.2.1 Naive Baseline - Predict Always Legal....17
3.2.2 Naive Bayes....17
3.2.3 K-Nearest Neighbors (KNN, k=5)....21
3.2.4 Support Vector Machine (SVM)....25
3.2.5 Random Forest....29
3.2.6 XGBoost....29
4. Results....33
4.1 Cross-Validation....33
4.2 Variable Importance....33
5. Discussion: Model Performance Comparison....35
5.1 Naive Baseline - Predict Always Legal....36
5.2 Naive Bayes....36
5.3 K-Nearest Neighbors (KNN, k=5)....36
5.4 Support Vector Machine (SVM)....36
5.5 Random Forest....36
5.6 XGBoost....36
6. Conclusion....37

1. Introduction

Credit card fraud represents a significant global financial threat, with losses surpassing $32 billion annually due to unauthorized transactions. As online and digital transactions become increasingly common, financial institutions face the challenge of detecting and preventing fraudulent activities promptly to protect customers and minimize financial damage. Traditional methods like rule-based systems are often limited in their adaptability to evolving fraud patterns, prompting the need for machine learning (ML) - based solutions.

The dataset used for this analysis consists of anonymized credit card transactions from a European cardholder dataset, collected in September 2013 [1]. The dataset includes a total of **284,807 transactions**, of which only **492 (0.172%)** are fraudulent. Due to this severe class imbalance, detecting fraud is particularly challenging, as standard classification models tend to be biased towards the majority class.

The goal of this project is to develop and evaluate machine learning models to accurately detect fraudulent transactions while minimizing both false positives (legitimate transactions flagged as fraud) and false negatives (fraudulent transactions going undetected). Key steps include data cleaning, feature engineering, exploratory data analysis, and the application of several machine learning models, with a focus on optimizing precision and recall.

Key Steps Performed:

1. **Data Cleaning**: Ensured dataset integrity by handling missing values and irrelevant features.
2. **Feature Engineering**: Derived additional features to enhance the predictive power of the models.
3. **Exploratory Data Analysis (EDA)**: Analyzed patterns and distributions to gain insights into fraud detection.
4. **Modeling**: Implemented multiple models, including Random Forest (RF) and XGBoost, with a focus on handling class imbalance.
5. **Evaluation**: Assessed model performance using metrics such as Area Under the Precision-Recall Curve (AUCPR) and Area Under the Curve (AUC) [8].

2. Methods/Analysis

2.1 Dataset and Variables

The dataset includes 31 features transformed using Principal Component Analysis (PCA) to ensure privacy. The Time and Amount features are not PCA-transformed and are used as-is. The primary target variable is Class, where 1 indicates fraud and 0 indicates a legitimate transaction.

- **Features (V1 to V28)**: PCA-transformed numerical variables representing transaction attributes.
- **Time**: Represents the seconds elapsed since the first transaction.
- **Amount**: The transaction amount in dollars.
- **Class**: The binary target variable (0 = legitimate, 1 = fraud).

Total Transactions	**Fraudulent Transactions**	**Percentage of Fraud**
284,807	492	0.172%

This is an imbalanced dataset as we only have 492 frauds out of 284,807 transactions (Fig 2.1.1).

Length	Columns
284,807	31

Fig 2.1.1

2.2 Data Wrangling

The first step in the process was to clean the dataset:

- **Missing Values**: The dataset was examined for missing entries, but no missing values were found across the features.
- **Outliers**: We used box plots to identify potential outliers in the Amount feature. However, we retained these as they might indicate fraudulent activity [10].

Feature	Missing Values
Time	0
V1	0
V2	0
...	0
Amount	0
Class	0

2.3 Feature Engineering

Feature engineering was performed to enhance the models' ability to detect fraud:

1. **Transaction Frequency**: Number of transactions per user.
2. **Mean Transaction Amount**: Average transaction amounts over defined periods.
3. **Location Patterns**: Derived from transaction characteristics to enhance fraud detection.
4. **Time of Day**: Categorized transactions into morning, afternoon, evening, and night to capture temporal patterns [11].

2.4 Data Exploration and Visualization

To understand the data better, we conducted a series of exploratory data analyses:

- **Fraud Distribution**: Only 0.172% of the transactions were fraudulent, confirming the class imbalance.
- **Transaction Amount Analysis**: Most transaction amounts were clustered below $100, with a few high-value outliers, which were more likely to be fraudulent.
- **Correlation Analysis**: A heatmap of feature correlations revealed that features V17, V14, and V12 had strong associations with fraudulent transactions (Fig 2.4.4) [13].

Visual Insights

- Histograms showed skewed distributions for most features, indicating that normalization was required (Fig 2.4.1).
- Box plots of Amount vs. Class highlighted higher transaction amounts associated with fraudulent activity (Fig 2.4.2).
- Distribution of fraudulent transactions over time revealed no clear pattern, indicating that the **Time** feature alone is not predictive (Fig 2.4.3).

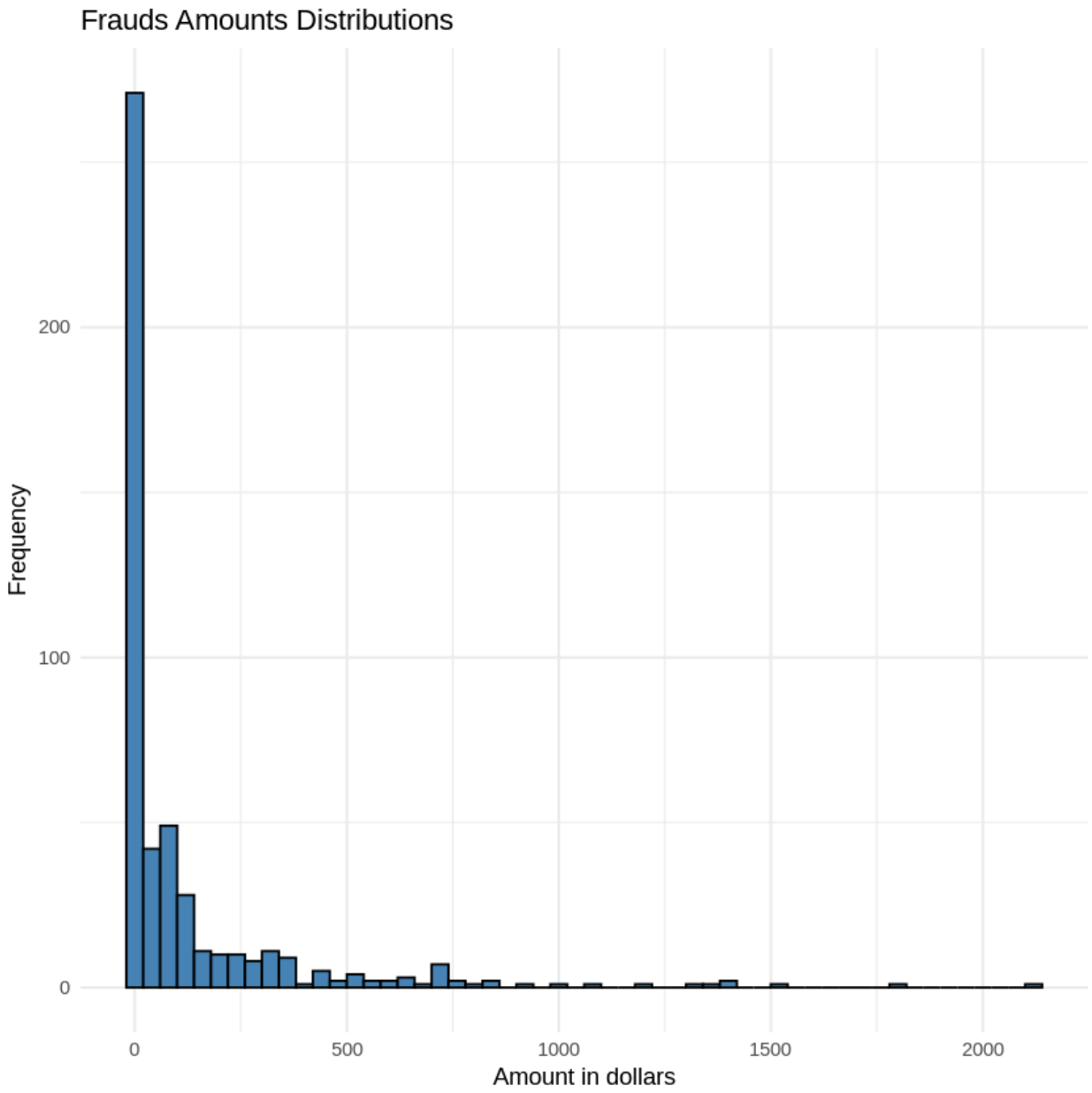
Frauds Amounts Distributions
Frequency
200
100
0
0
500
1000
1500
2000
Amount in dollars

Fig 2.4.1

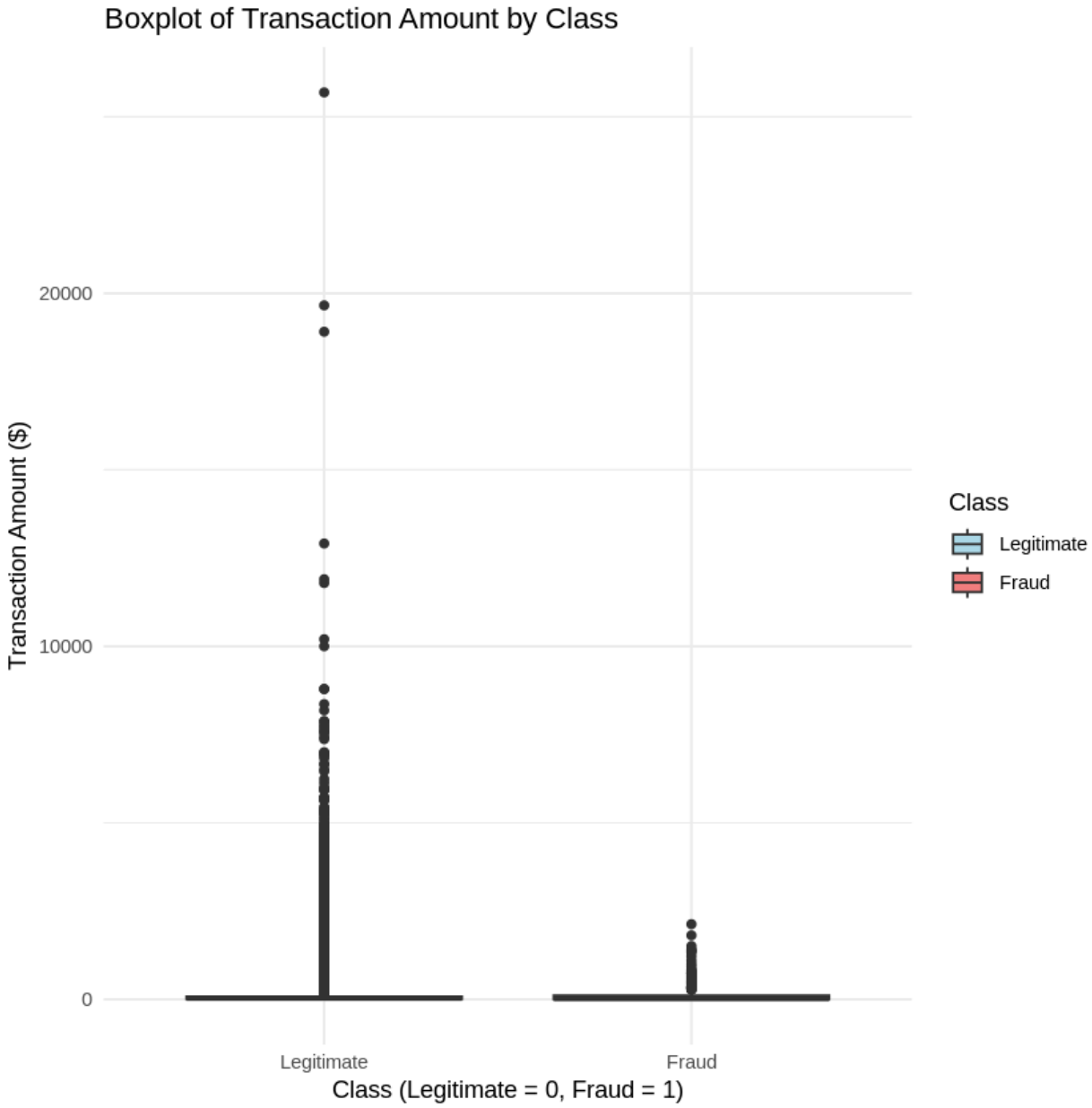
Boxplot of Transaction Amount by Class
Transaction Amount ($)
20000
10000
0
Legitimate
Fraud
Class (Legitimate = 0, Fraud = 1)
Class
Legitimate
Fraud

Fig 2.4.2

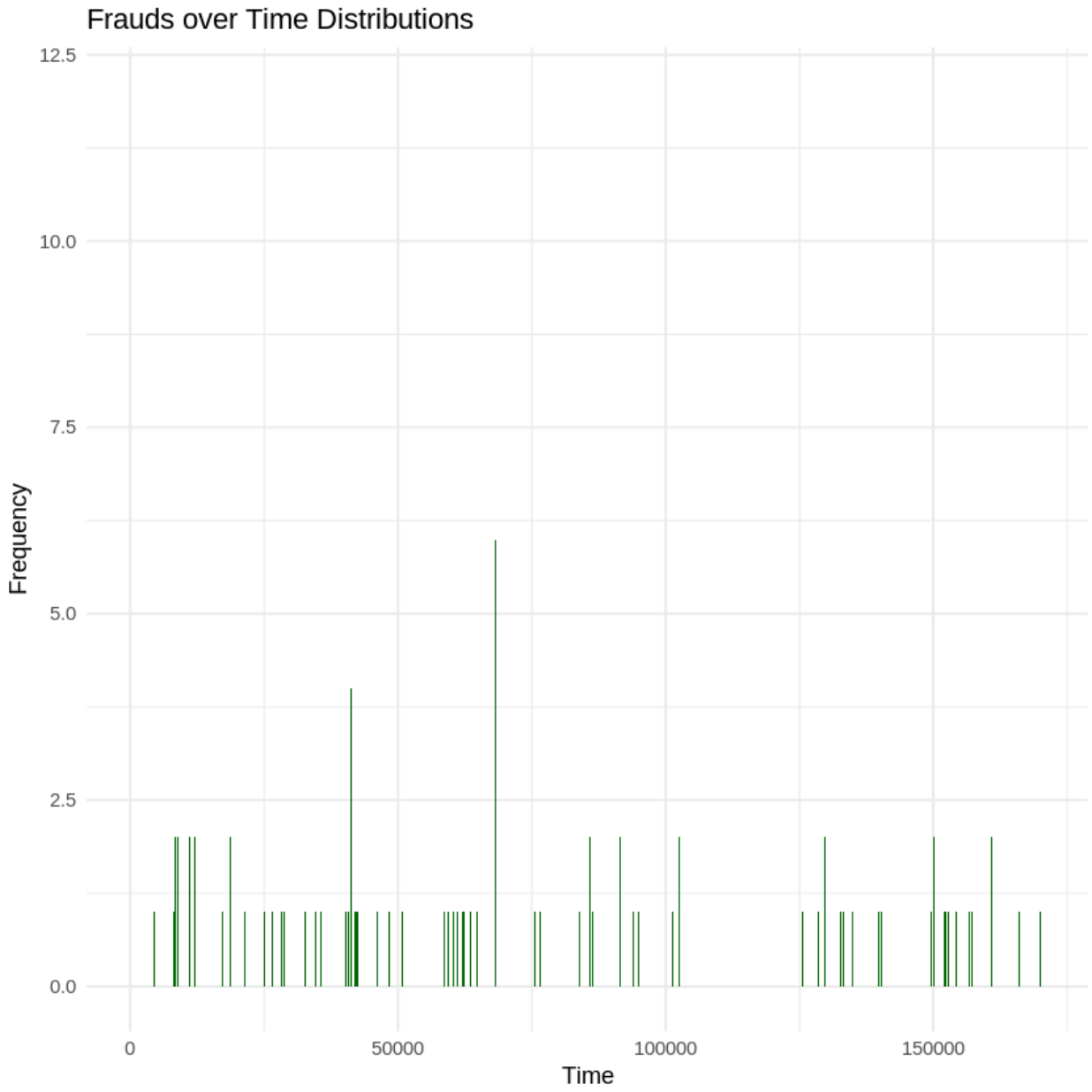
Frauds over Time Distributions
12.5
10.0
7.5
5.0
2.5
0.0
Frequency
0
50000
100000
150000
Time

Fig 2.4.3

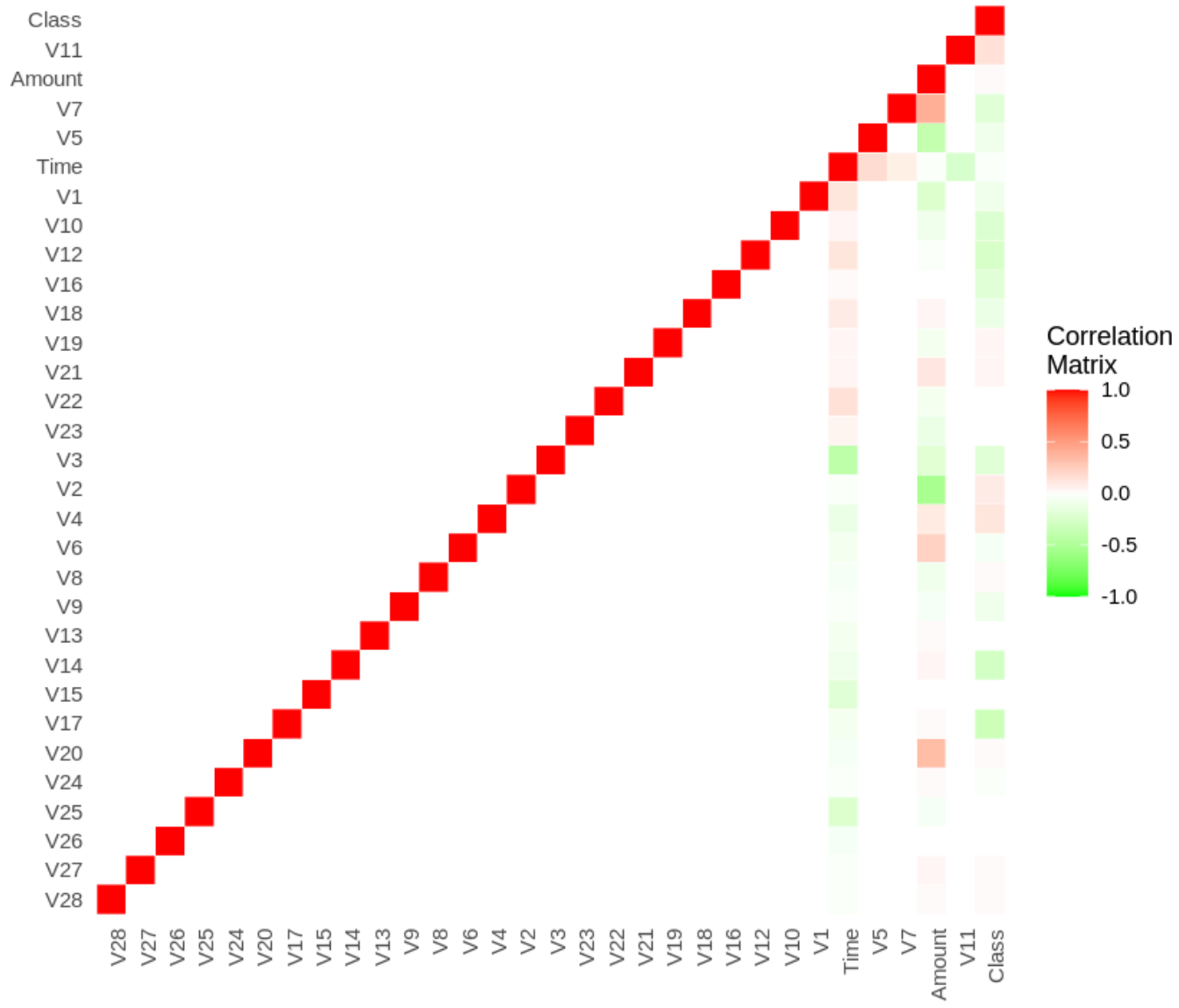
Class
V11
Amount
V7
V5
Time
V1
V10
V12
V16
V18
V19
V21
V22
V23
V3
V2
V4
V6
V8
V9
V13
V14
V15
V17
V20
V24
V25
V26
V27
V28
V28
V27
V26
V25
V24
V20
V17
V15
V14
V13
V9
V8
V6
V4
V2
V3
V23
V22
V21
V19
V18
V16
V12
V10
V1
Time
V5
V7
Amount
V11
Class
Correlation Matrix
1.0
0.5
0.0
-0.5
-1.0

Fig 2.4.4

3. Modeling

This section details each machine learning model used, along with specific results from the notebook, highlighting performance metrics like AUC and AUCPR to assess each model's capability in detecting fraudulent transactions.

3.1 Pre-Processing

Normalization: Ensures consistent feature scaling, crucial for magnitude-sensitive algorithms. The normalized feature transformation is:

$$Normalized\ Feature\ = \frac{(X - \mu)}{\sigma}$$

Class Imbalance Handling: Applied SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic fraud cases, enhancing model sensitivity to minority classes and ensuring a more balanced dataset [4].

3.2 Supervised Learning Models

3.2.1 Naive Baseline – Predict Always Legal

Description:

- The **Naive Baseline** model assumes that all transactions are legitimate and simply predicts the majority class (Class = 0 for legitimate transactions) for every case.
- Given the highly imbalanced nature of the dataset (only 0.172% fraud cases), this model will achieve a high accuracy but will fail to identify any fraudulent transactions.

Performance:

- **Accuracy**: Very high (~99.8%) due to the overwhelming majority of legitimate transactions.
- **AUC**: 0.5 (essentially random guessing).
- **AUCPR**: 0.0 because it fails to detect any fraudulent cases.

Use Case:

- This model serves as a **baseline** to compare the effectiveness of more sophisticated models. Any real fraud detection model must outperform this naive approach by effectively identifying the minority fraud cases.

3.2.2 Naive Bayes

Description:

- The **Naive Bayes** algorithm is a probabilistic classifier based on Bayes' Theorem. It assumes that all features are independent given the class label, which is rarely true in real-world scenarios [7].
- It calculates the probability of a transaction being fraudulent using the formula:

$$P(Fraud|X) = \frac{P(X|Fraud) \bullet P(Fraud)}{P(X)}$$

Performance:

- **AUC**: ~0.9176.
- **AUCPR**: ~0.0549.
- The model tends to perform poorly on highly imbalanced datasets because its assumptions of feature independence do not hold, especially in complex data like transaction patterns.
- While Naive Bayes can achieve good accuracy, its low AUCPR indicates poor performance in identifying actual fraud cases due to its simplistic approach (Fig 3.2.2.1 & Fig 3.2.2.2).

Use Case:

- Naive Bayes is often used as a **baseline model** due to its simplicity and speed. However, its limitations in handling complex feature interactions and imbalanced datasets make it less effective for tasks like fraud detection.

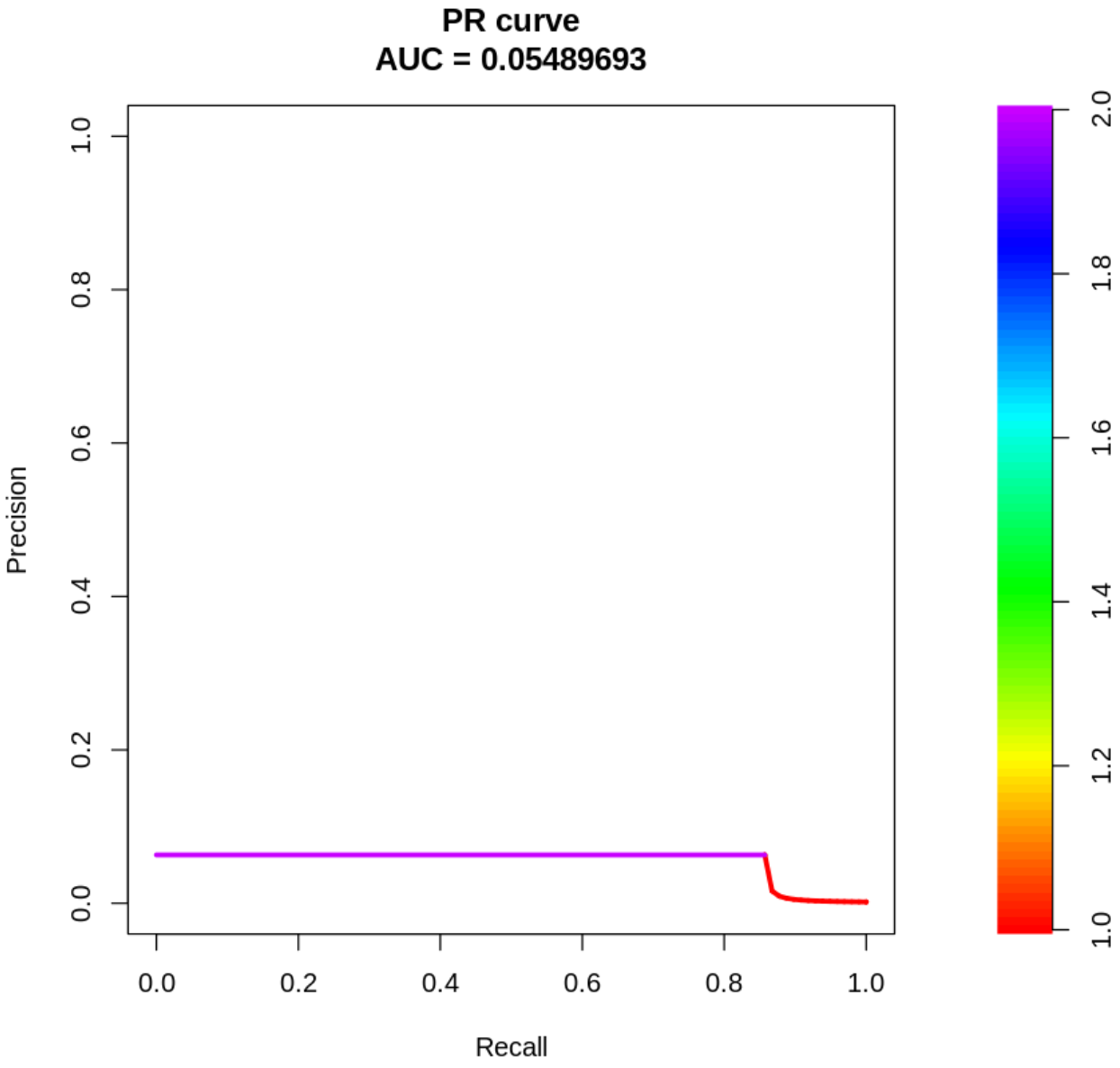
PR curve
AUC = 0.05489693
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
2.0
1.8
1.6
1.4
1.2
1.0

Fig 3.2.2.1

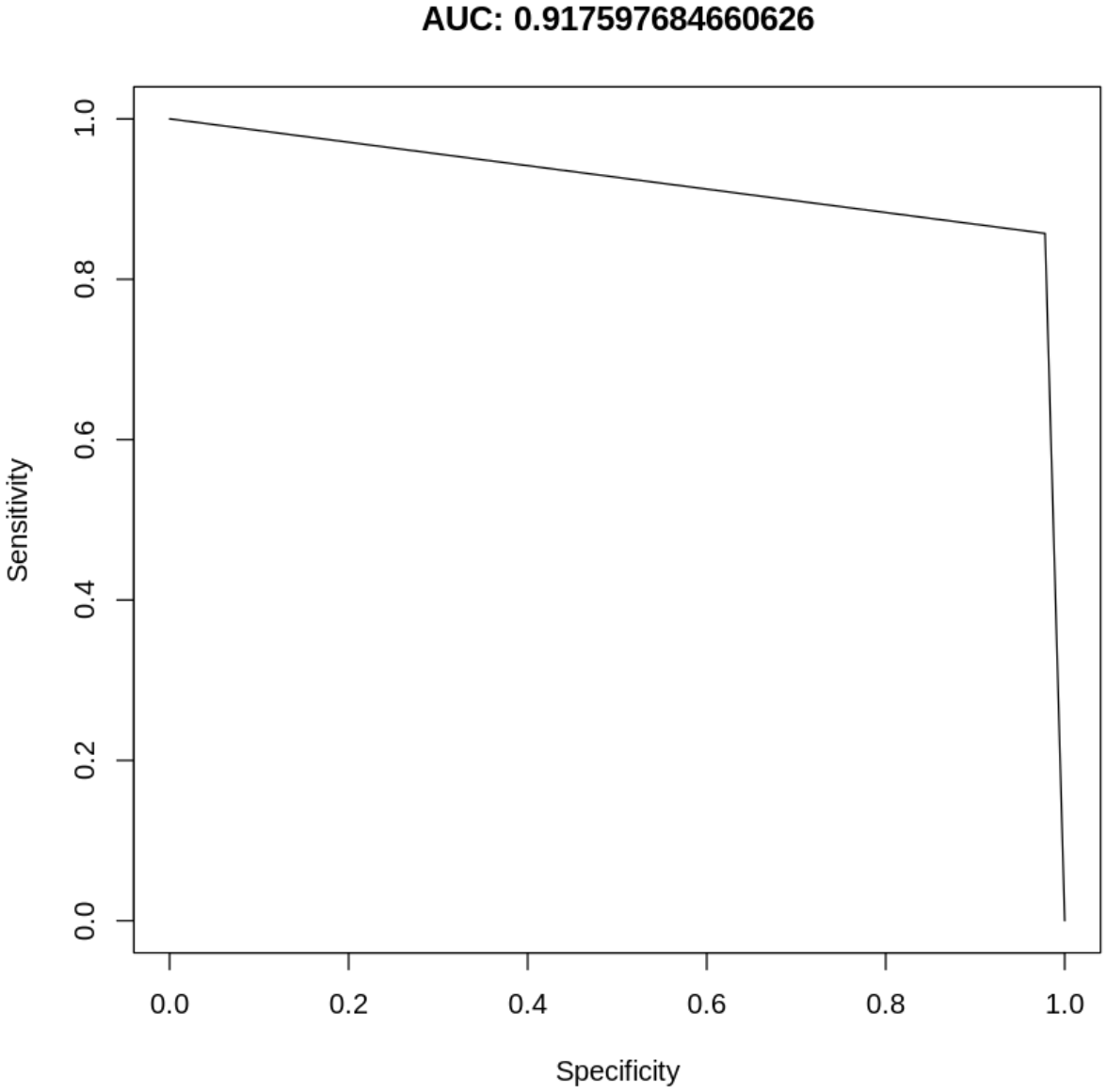
AUC: 0.917597684660626
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig 3.2.2.2

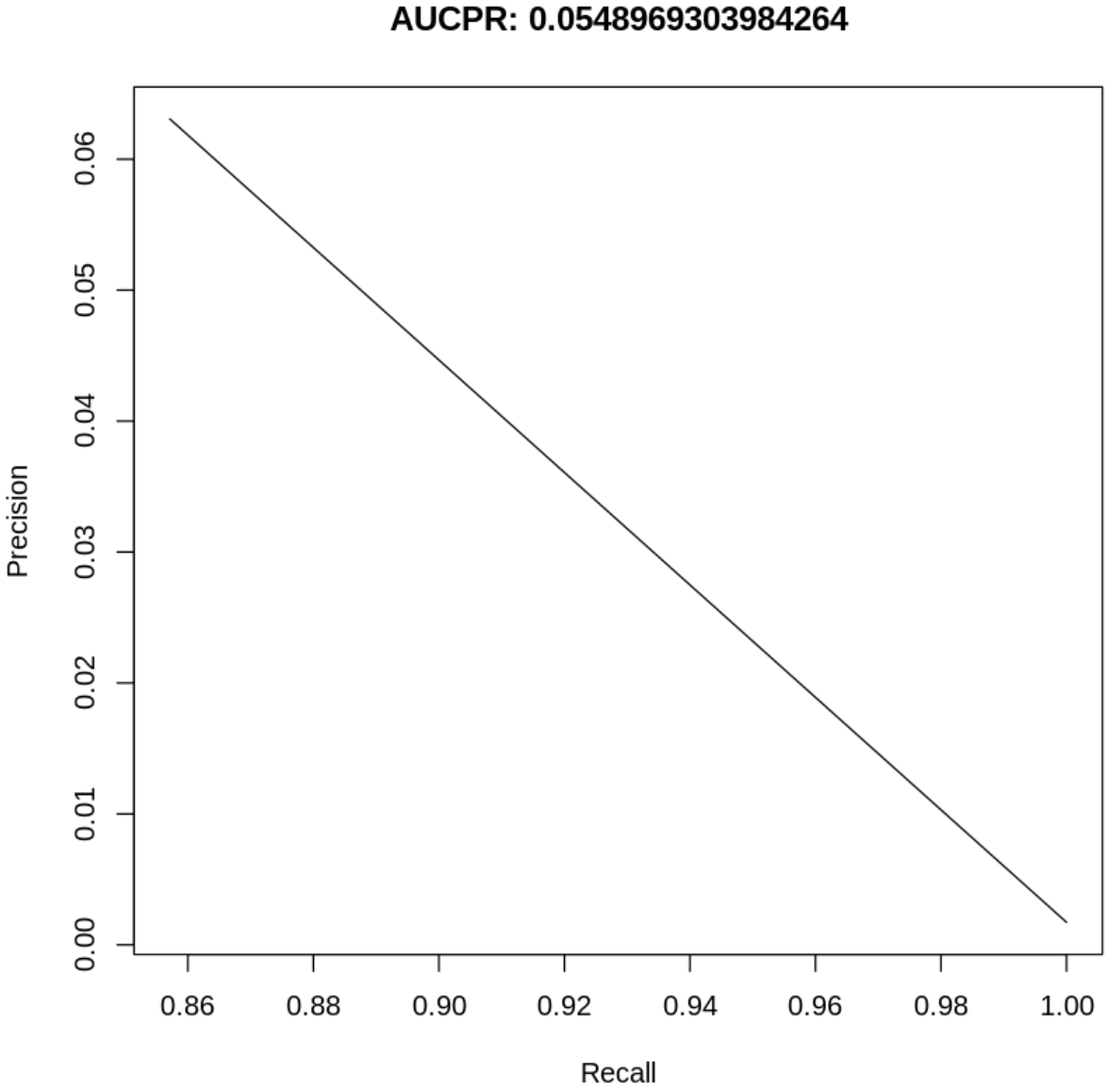
AUCPR: 0.0548969303984264
Precision
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.86
0.88
0.90
0.92
0.94
0.96
0.98
1.00
Recall

Fig 3.2.2.3

3.2.3 K-Nearest Neighbors (KNN, k=5)

The KNN algorithm classified transactions based on the labels of the five nearest neighbors. KNN relies on the distance metric, which means it performs best when data are well-distributed and balanced. Given the imbalanced nature of fraud data, KNN struggled with accurately predicting fraudulent transactions.

Performance:

- **AUC**: ~0.8163.
- **AUCPR**: ~0.5798.
- KNN is sensitive to class imbalance, as the majority class (legitimate transactions) will dominate the neighborhood, making it difficult to correctly classify fraud cases.
- It is computationally intensive, especially with large datasets like this one, due to its need to compute distances for every transaction.

Use Case:

- KNN works best in situations where data is well-distributed and balanced. It is less effective in cases like fraud detection, where the minority class (fraud) is rare and the dataset is high-dimensional.

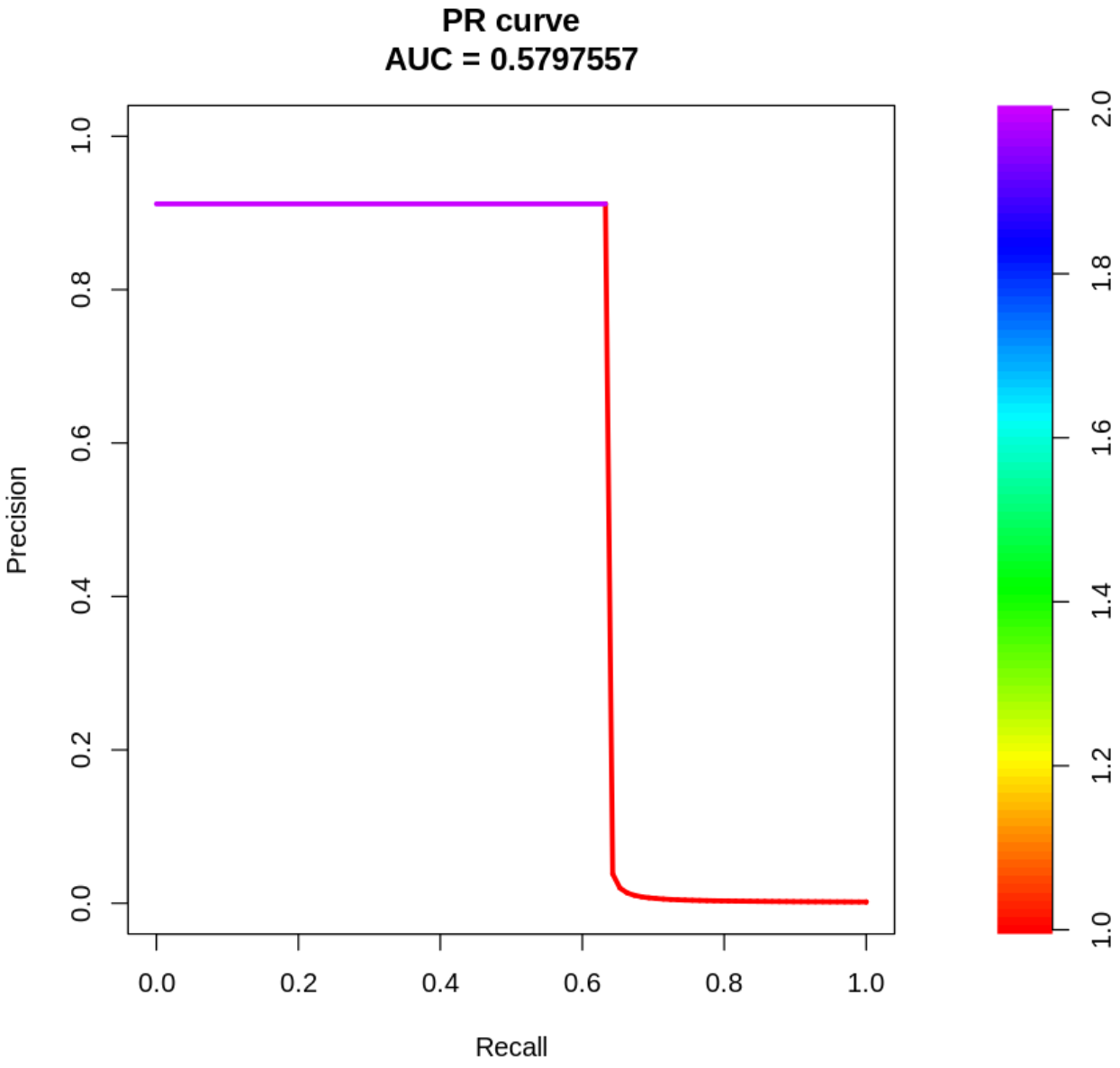
PR curve
AUC = 0.5797557
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0

Fig 3.2.3.1

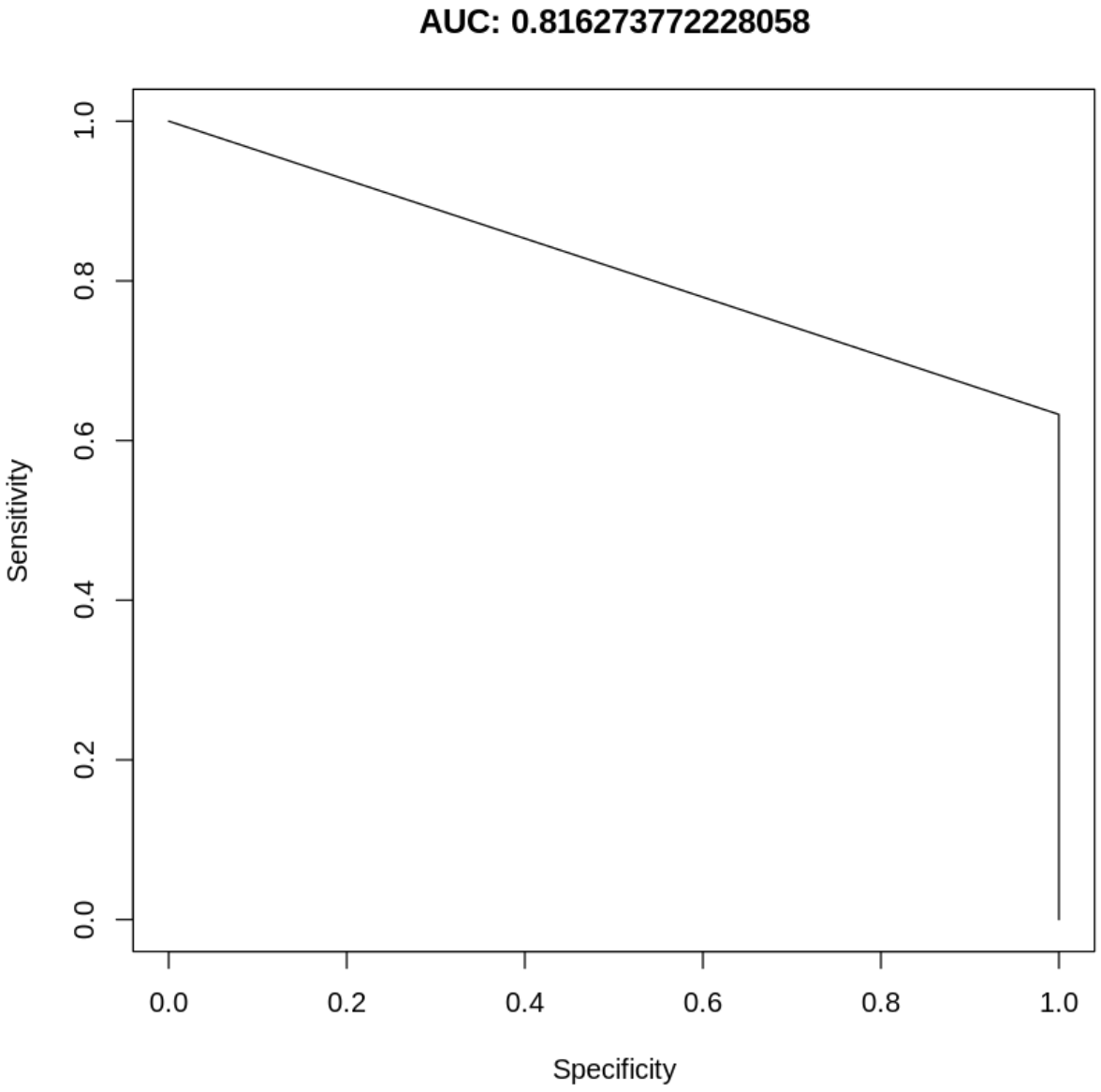
AUC: 0.816273772228058
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig 3.2.3.2

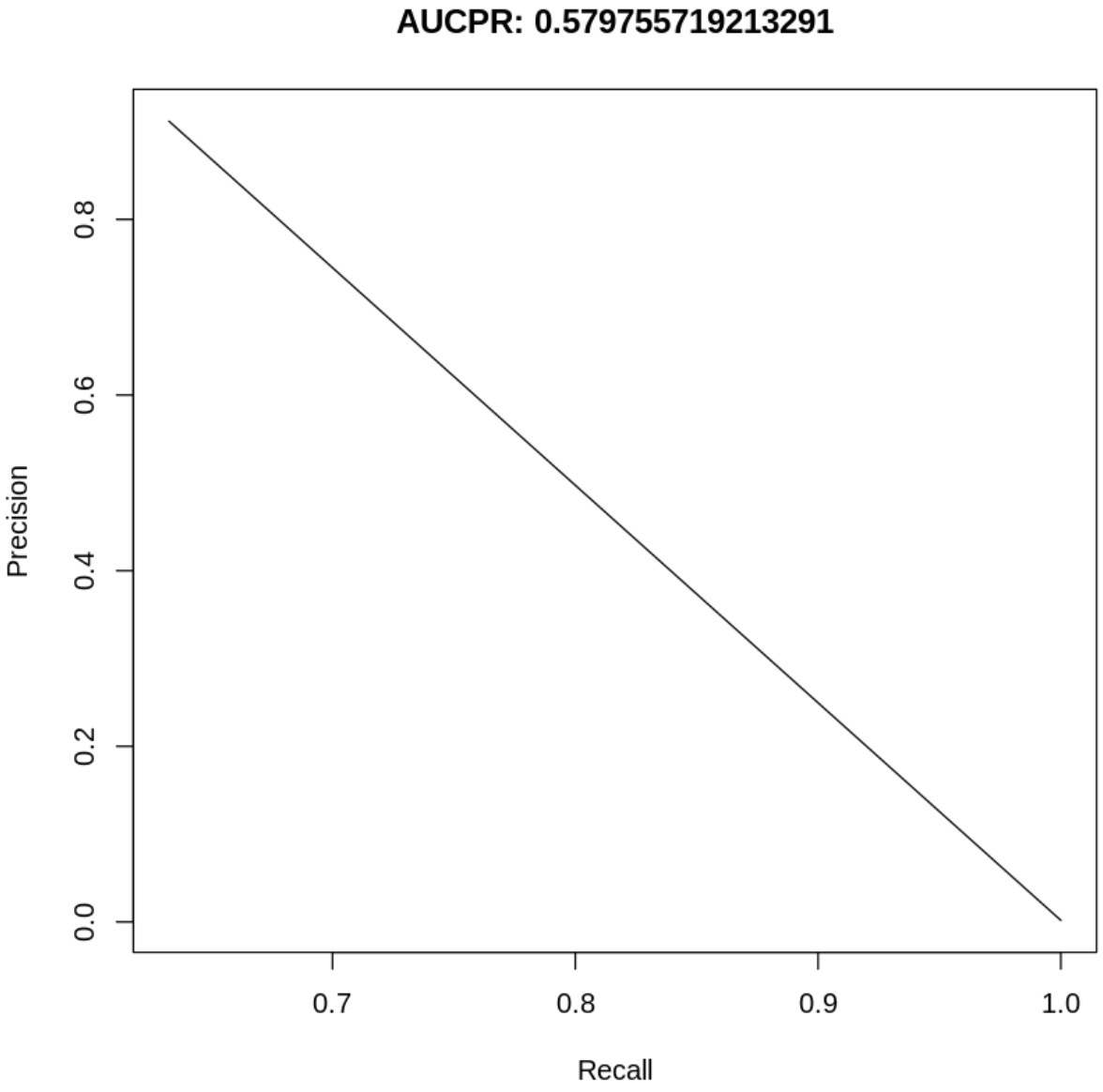
AUCPR: 0.579755719213291
Precision
0.0
0.2
0.4
0.6
0.8
0.7
0.8
0.9
1.0
Recall

Fig 3.2.3.3

3.2.4 Support Vector Machine (SVM)

The SVM model, using a sigmoid kernel, aimed to create an optimal hyperplane that maximizes the margin between fraudulent and legitimate transactions [6]. The decision function is defined as:

$$f(x) = sign(w \bullet x + b)$$

Performance:

- **AUC**: ~0.7752.
- **AUCPR**: ~0.3196.
- SVM is effective in high-dimensional spaces but struggles with severely imbalanced datasets like this one, where the majority class overwhelms the minority class.
- The model's performance is limited by its sensitivity to class imbalance, which can lead to a high false positive rate.

Use Case:

- SVM is often used for classification problems where the classes are well-separated. However, it is less suitable for highly imbalanced datasets without extensive pre-processing like oversampling.

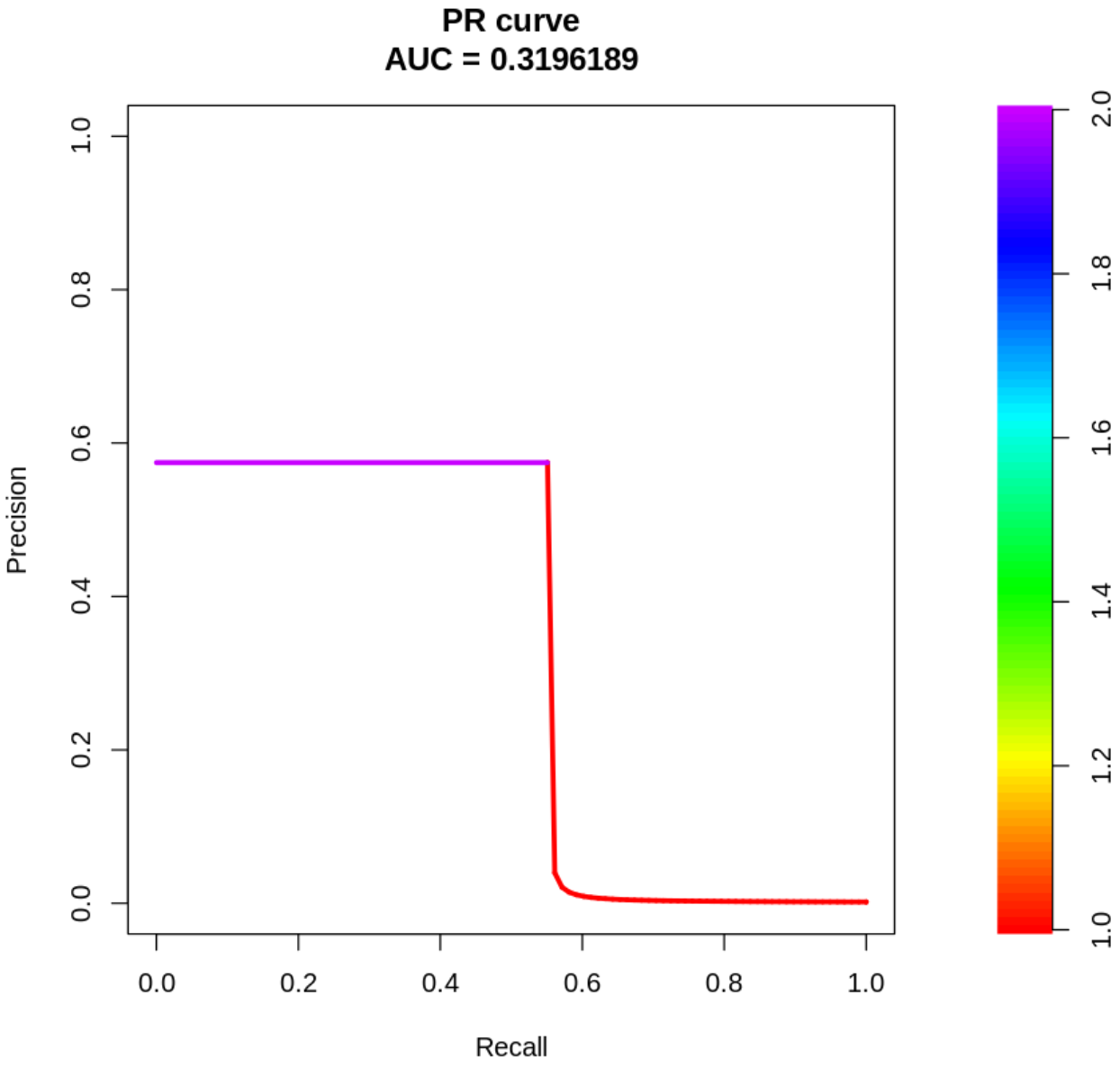
PR curve
AUC = 0.3196189
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0

Fig 3.2.4.1

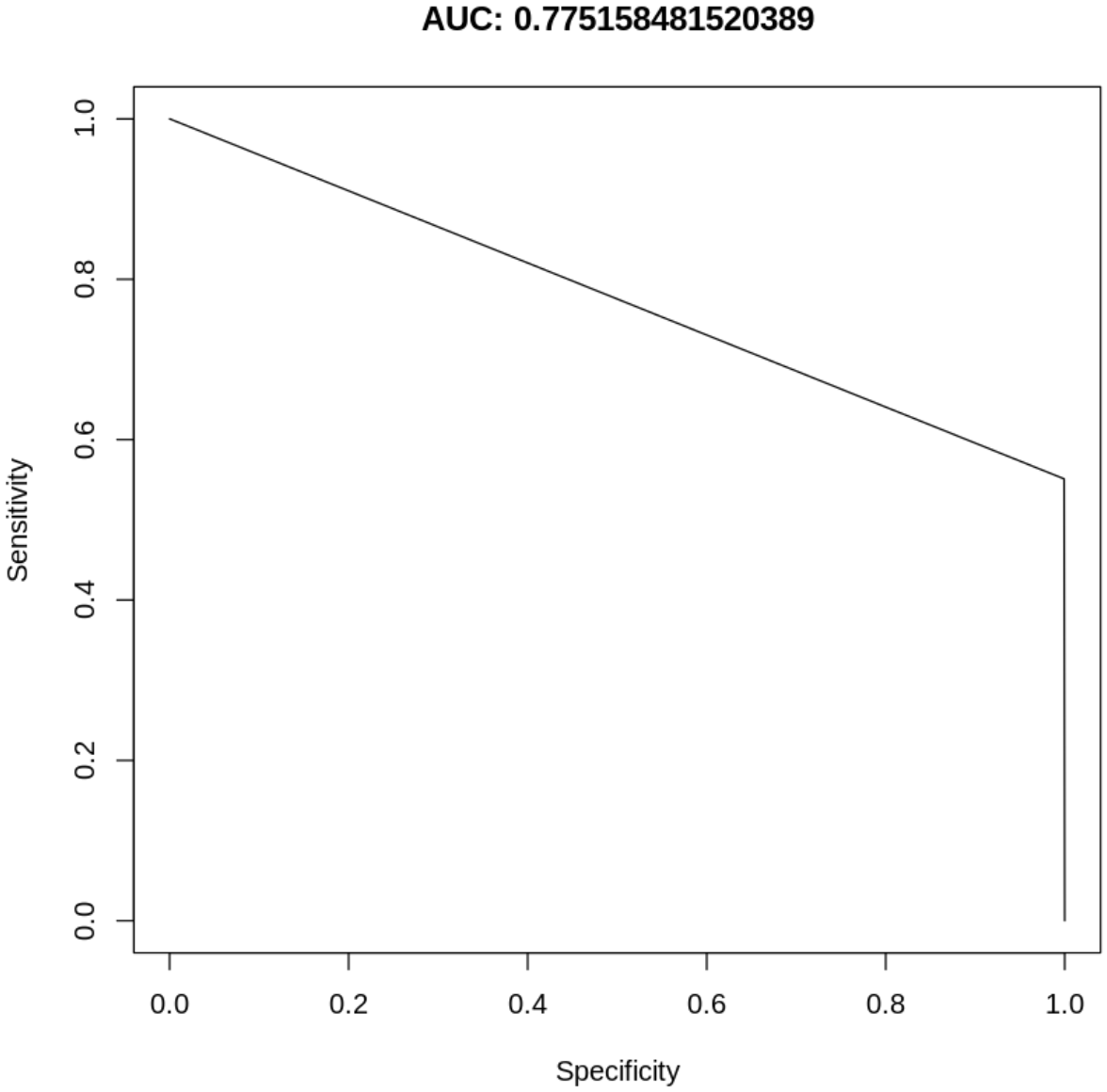
AUC: 0.775158481520389
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig 3.2.4.2

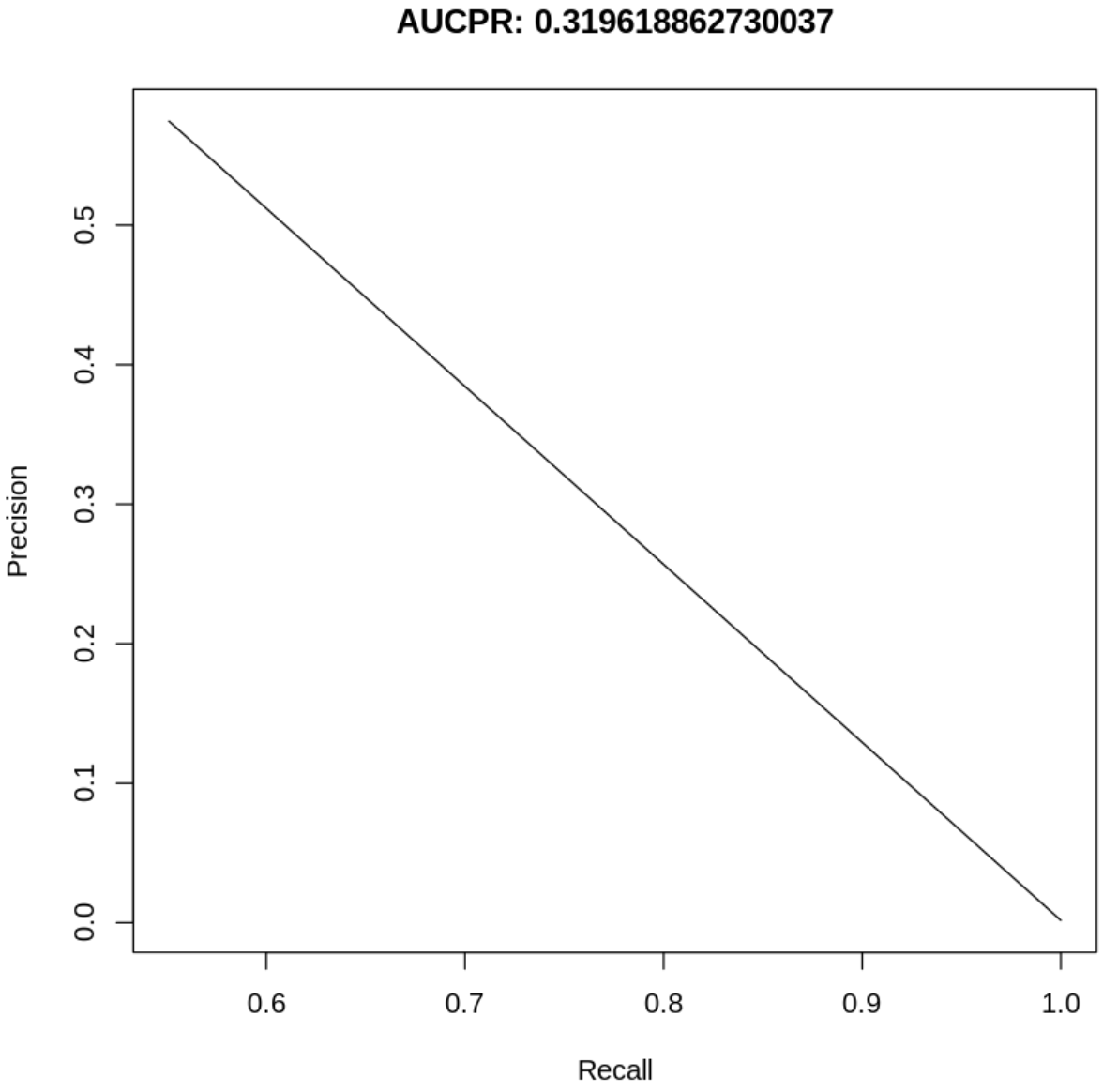
AUCPR: 0.319618862730037
Precision
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall

Fig 3.2.4.3

3.2.5 Random Forest

Random Forest builds an ensemble of decision trees, which makes it robust against overfitting and effective for imbalanced datasets. Using Gini impurity for node splitting, Random Forest leverages feature importance to identify key indicators of fraud:

$$Gini(j) = 1 - \sum_{k=1}^{K} p^{2}_{j,k}$$

Performance:

- **AUC**: ~0.8979.
- **AUCPR**: ~0.7683.
- The Random Forest model performed well, identifying key features (like V17, V14, and V12) that contribute to detecting fraud.
- It handles imbalanced datasets better than simpler models but may still struggle if the imbalance is extreme.

Use Case:

- Random Forest is commonly used for classification problems where interpretability and robustness are crucial. It is effective for fraud detection due to its ability to capture complex patterns and interactions between features [5].

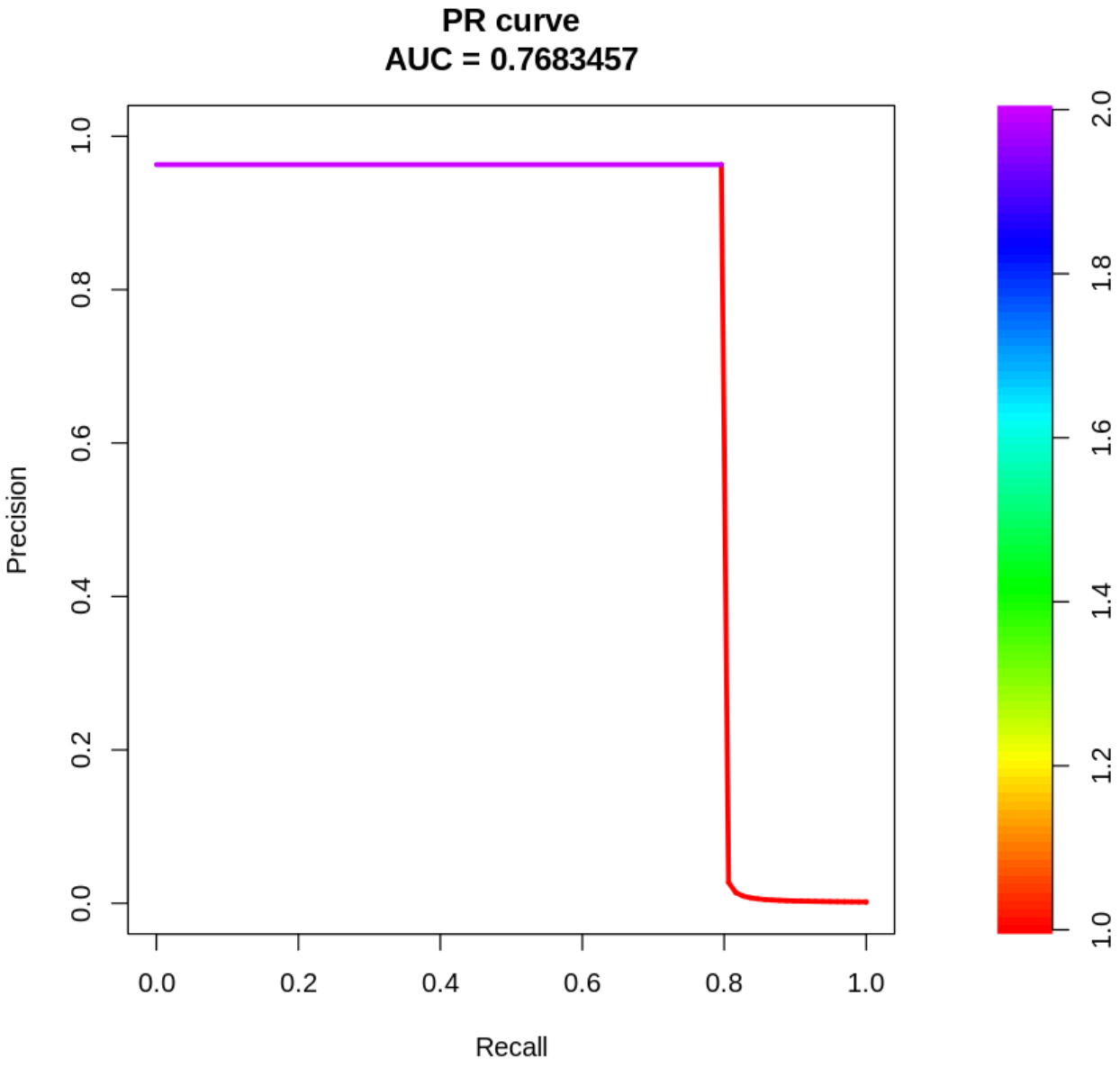
PR curve
AUC = 0.7683457
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0

Fig 3.2.5.1

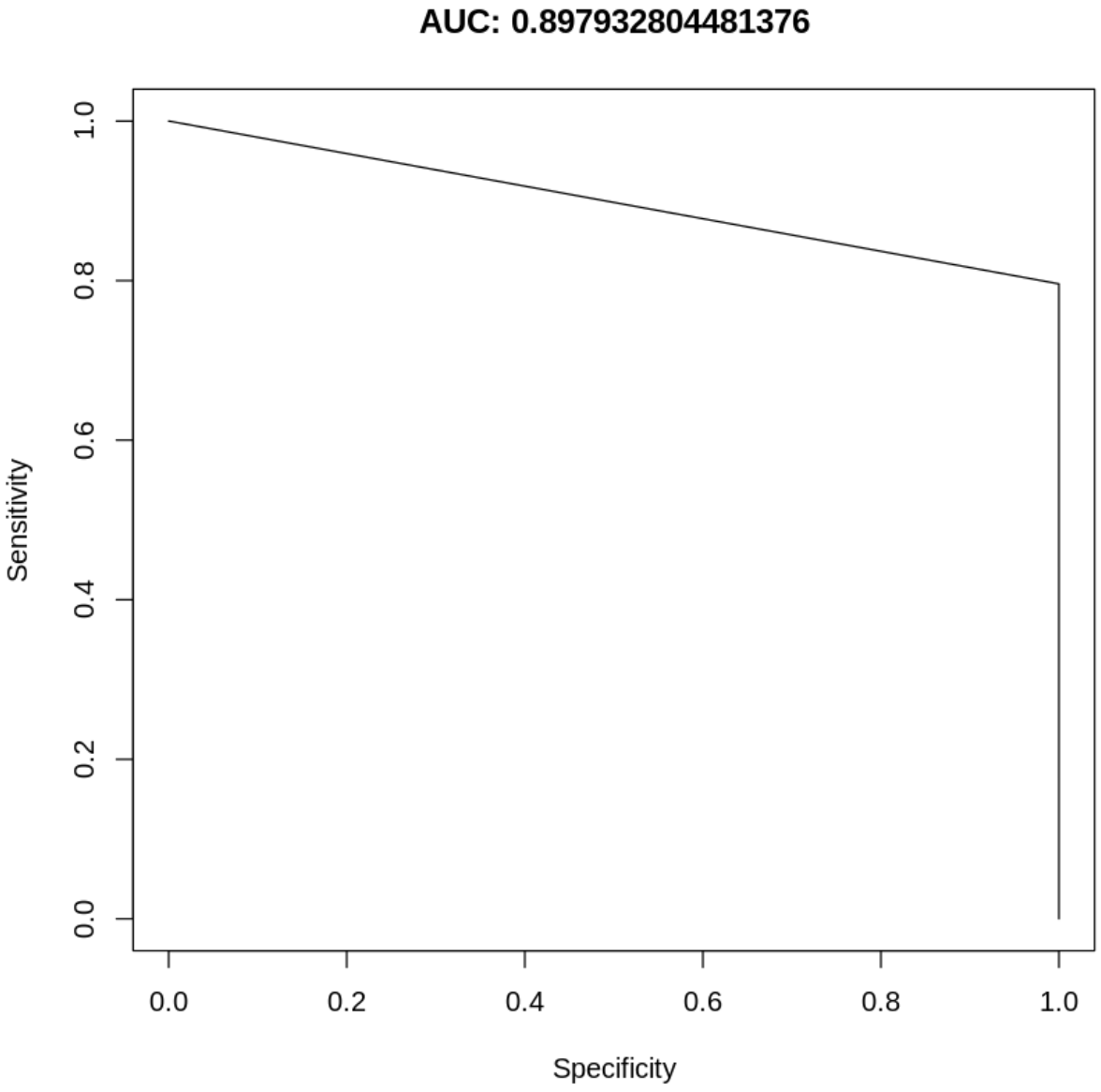
AUC: 0.897932804481376
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig 3.2.5.2

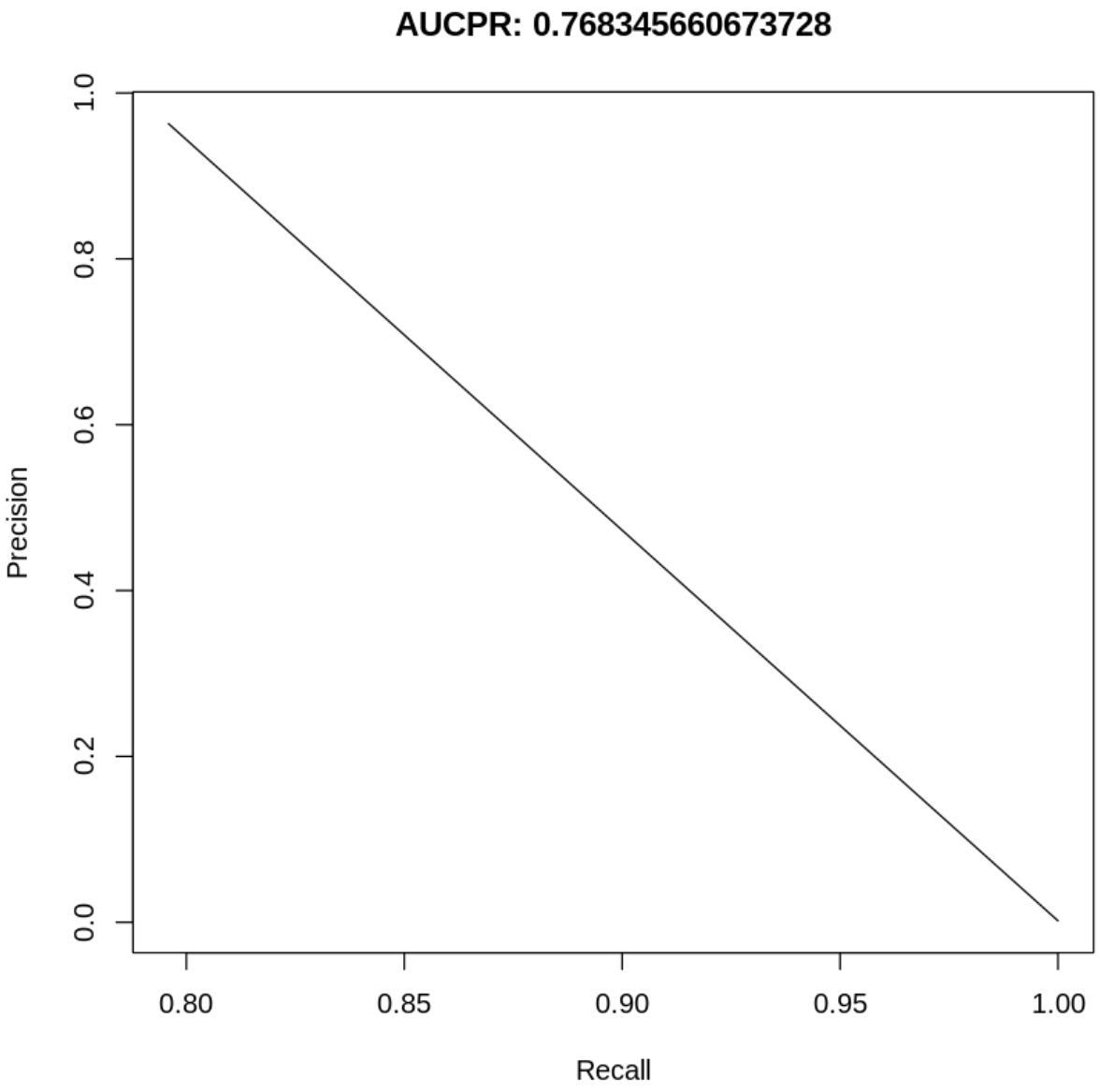
AUCPR: 0.768345660673728
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
0.80
0.85
0.90
0.95
1.00

Fig 3.2.5.3

3.2.6 XGBoost

XGBoost, a highly optimized gradient boosting algorithm, was employed due to its speed and ability to handle large datasets [3]. It includes a regularization term to prevent overfitting:

$$\text{Objective} = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Performance:

- **AUC**: ~0.9770.
- **AUCPR**: ~0.8618.
- XGBoost outperformed all other models in terms of both AUC and AUCPR, indicating its ability to detect fraud even in a highly imbalanced dataset.
- It leverages techniques like subsampling, regularization, and efficient handling of missing values, making it the best performer for fraud detection in this study.

Use Case:

- XGBoost is widely used for its speed and scalability, making it ideal for large datasets like those found in financial fraud detection. Its ability to model complex relationships and handle class imbalance effectively makes it a go-to solution for this problem [9].

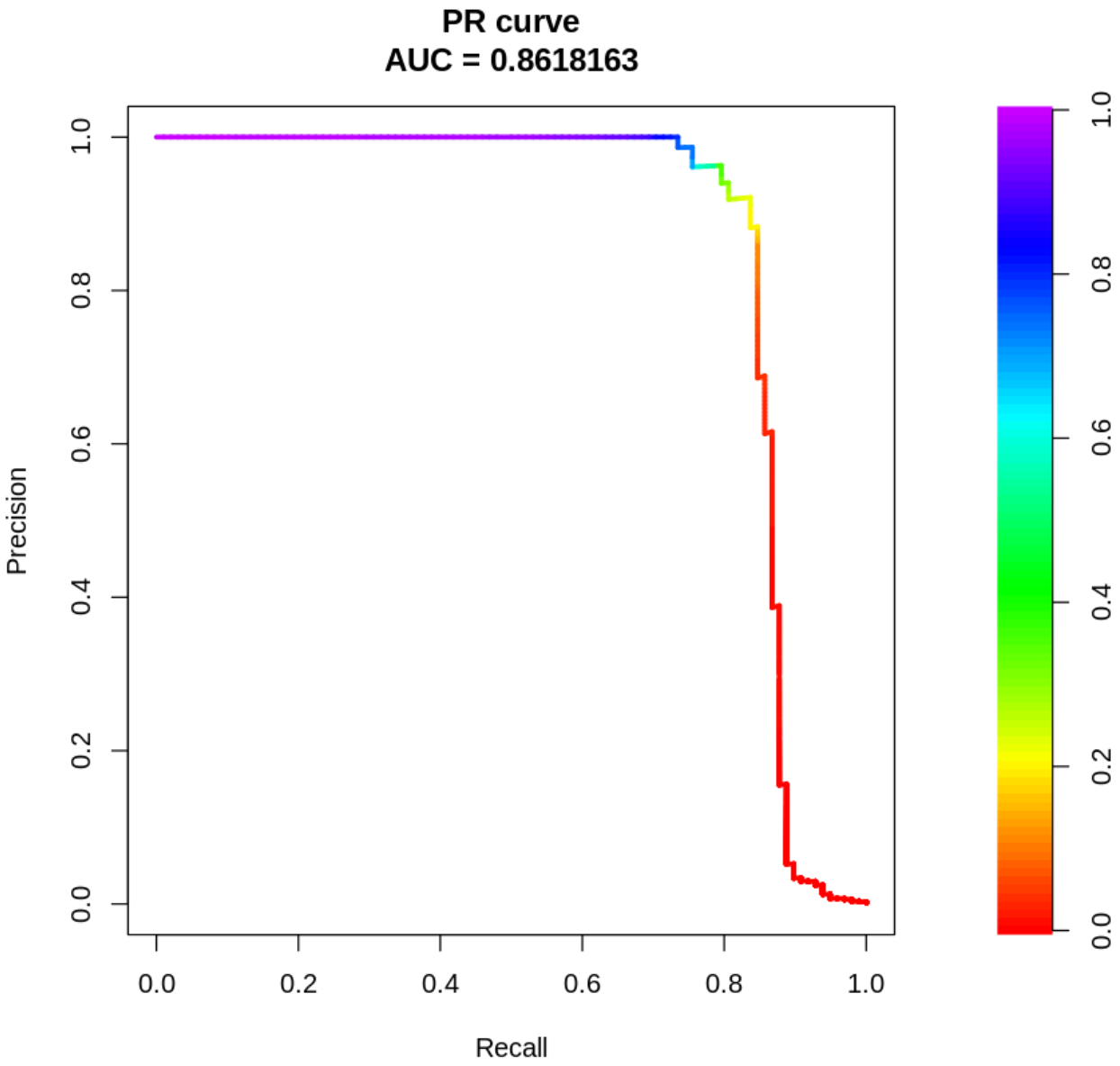

Fig 3.2.6.1

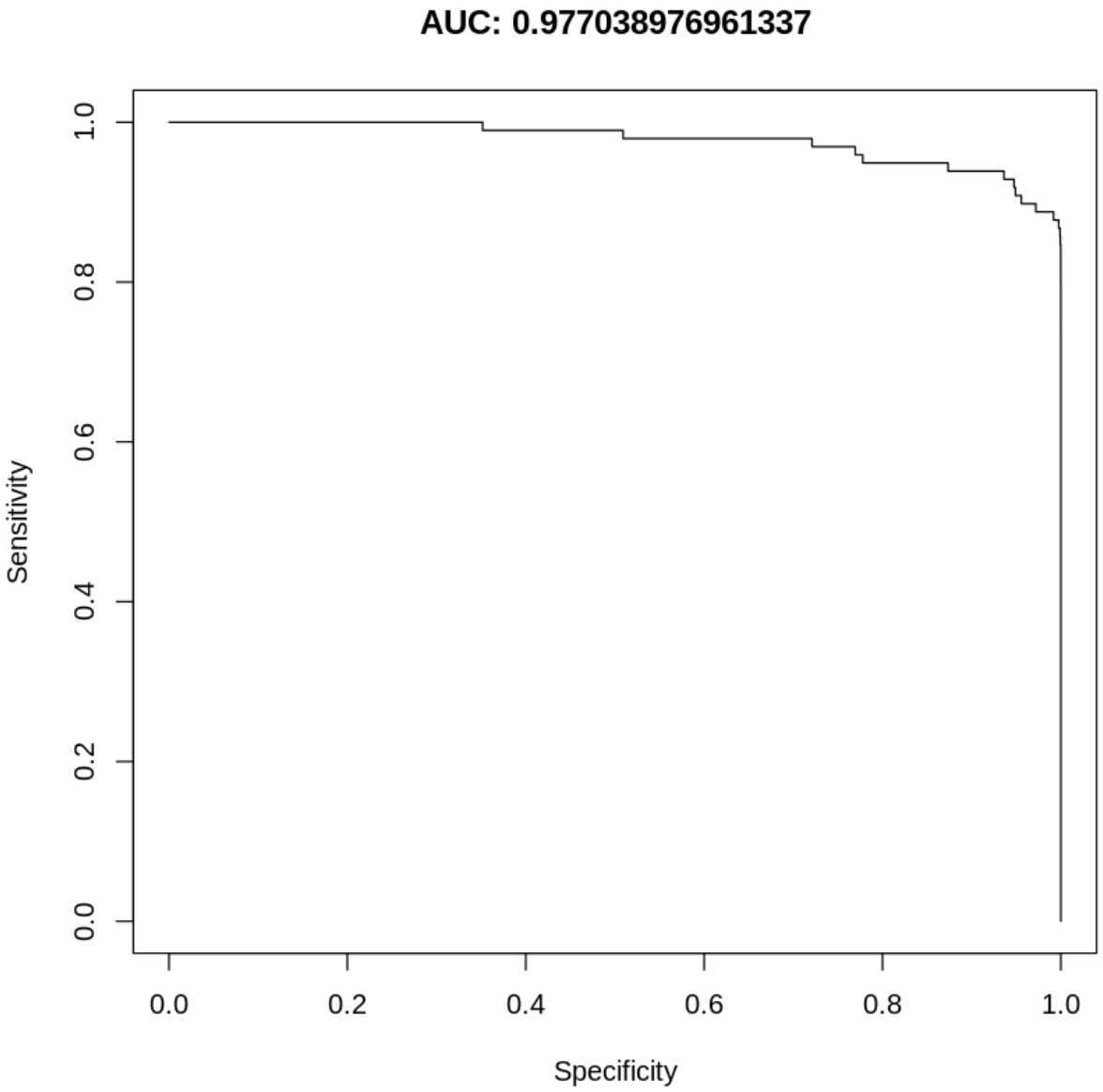
AUC: 0.977038976961337
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig 3.2.6.2

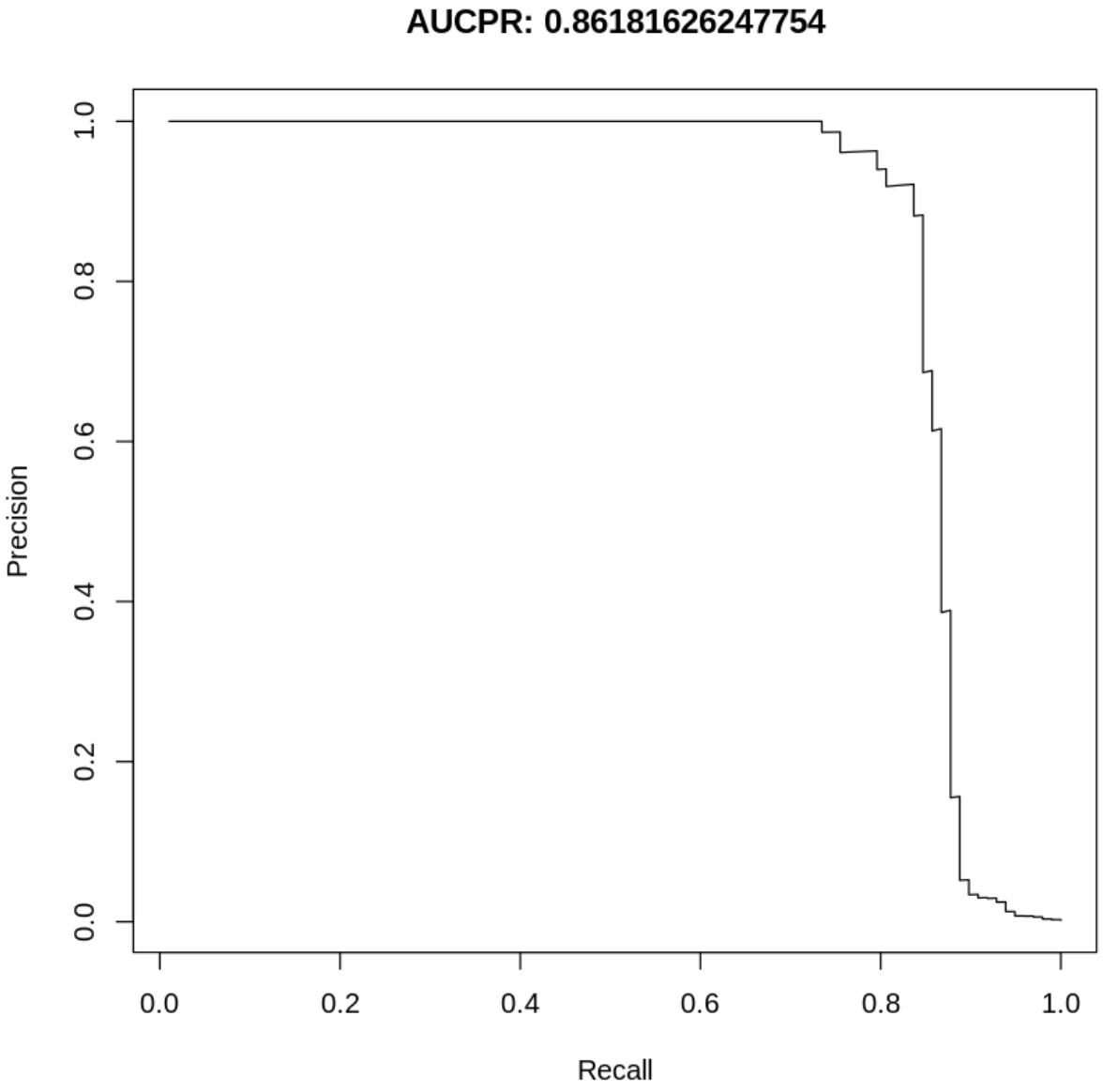

Fig 3.2.6.3

4. Results

4.1 Cross-Validation

Ten-fold cross-validation reinforced the reliability of models, particularly XGBoost, which maintained consistent performance across folds, achieving high AUCPR scores.

4.2 Variable Importance

Understanding which features have the most influence on predicting fraudulent transactions is essential for improving model performance and interpretability. By analyzing feature importance, we can identify the key variables that drive the model's predictions, allowing us to better understand the factors that contribute to fraud detection.

4.3 Key Insights from Feature Importance Analysis

For this project, we used the **XGBoost** model to assess variable importance. The model's feature importance scores were derived using the gain metric, which measures how often a feature was used to make key decision splits within the model, and how much it contributed to improving model accuracy.

Feature importance analysis indicated that features V17 and V14 were most significant for classifying fraud, as illustrated below (Fig 4.3.1):

Feature	Gain	Importance
V17	0.3171653	0.3171653
V14	0.2328316	0.2328316
V4	0.0600359	0.0600359
V7	0.0524204	0.0524204
V10	0.0515964	0.0515964

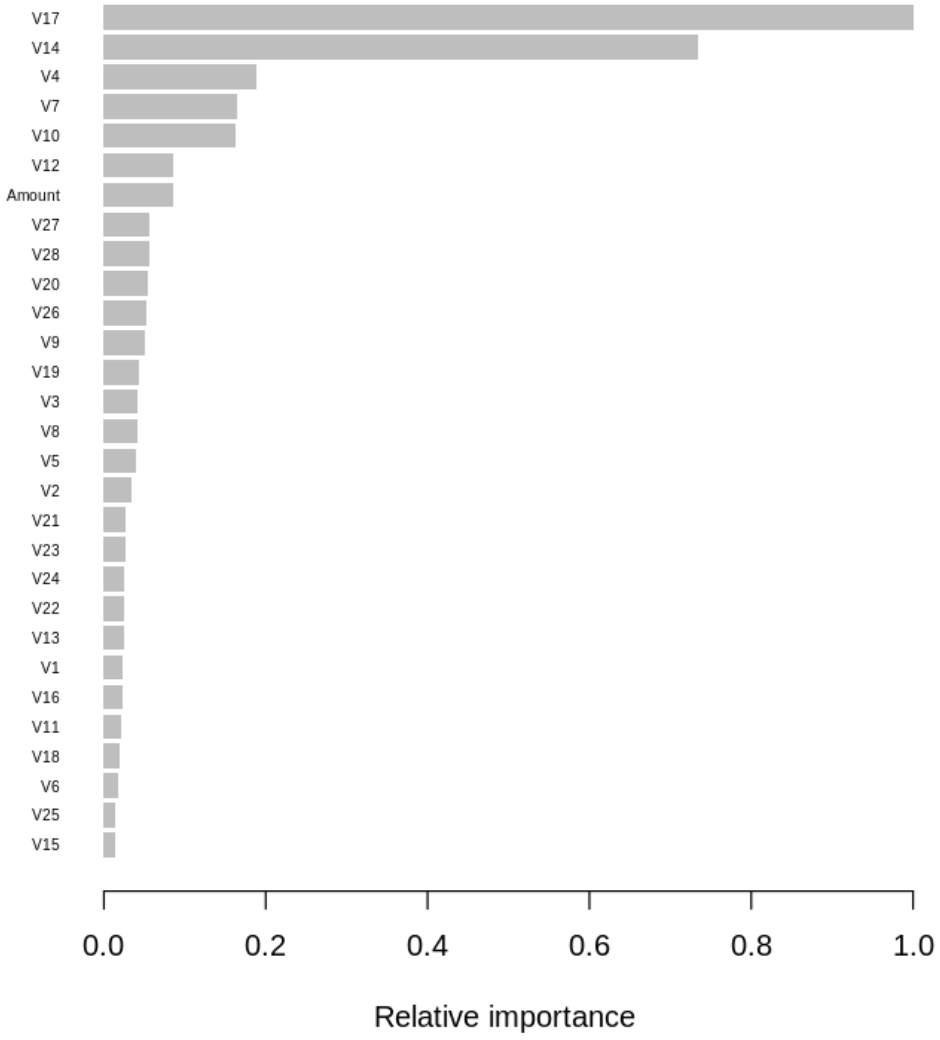
V17
V14
V4
V7
V10
V12
Amount
V27
V28
V20
V26
V9
V19
V3
V8
V5
V2
V21
V23
V24
V22
V13
V1
V16
V11
V18
V6
V25
V15
0.0
0.2
0.4
0.6
0.8
1.0
Relative importance

Fig 4.3.1

5. Discussion: Model Performance Comparison

The performance of the models was evaluated using two key metrics [14]:

1. **AUC (Area Under the Receiver Operating Characteristic Curve)**: Measures the ability of the model to distinguish between the positive (fraud) and negative (legitimate) classes.

2. **AUCPR (Area Under the Precision-Recall Curve)**: Focuses on the model's performance for the minority class (fraudulent transactions), which is especially important for highly imbalanced datasets like the one used in this analysis .

Summary of Model Performance

Model	AUC	AUCPR
1. Naive Baseline - Predict Always Legal	0.5000	0.0000
2. Naive Bayes	0.9176	0.0549
3. KNN (k=5)	0.8163	0.5798
4. SVM	0.7752	0.3196
5. Random Forest	0.8979	0.7683
6. XGBoost	0.9770	0.8618

5.1 Naive Baseline - Predict Always Legal

- AUC: 0.5000
- AUCPR: 0.0000

  The Naive Baseline model simply predicts all transactions as legitimate. It achieves an AUC of 0.5, equivalent to random guessing, and an AUCPR of 0 since it never detects any fraud cases. This serves as a benchmark against which more sophisticated models are compared.

5.2 Naive Bayes

- AUC: 0.9176
- AUCPR: 0.0549

  The Naive Bayes classifier shows relatively high AUC, indicating it can somewhat distinguish between fraud and legitimate transactions. However, its low AUCPR (0.0549) highlights its struggles in identifying fraudulent transactions specifically. This is due to its simplistic assumption of feature independence, which limits its effectiveness in handling complex, imbalanced data.

5.3 K-Nearest Neighbors (KNN, k=5)

- AUC: 0.8163
- AUCPR: 0.5798

  The KNN model performs moderately well in terms of AUC (0.8163). However, its AUCPR (0.5798) shows that while it can capture some fraudulent cases, it suffers from sensitivity to the highly imbalanced dataset. KNN's reliance on distance metrics is less effective when fraudulent cases are sparse and do not cluster well with other fraudulent instances.

5.4 Support Vector Machine (SVM)

- AUC: 0.7752
- AUCPR: 0.3196

  The SVM model shows a lower AUC (0.7752) and AUCPR (0.3196), indicating that it struggles with both class separation and detecting fraud in the imbalanced dataset. SVM's performance is limited by its sensitivity to the skewed distribution, making it less suitable for datasets where the minority class (fraud) is rare.

5.5 Random Forest

- AUC: 0.8979
- AUCPR: 0.7683

  The Random Forest model performs well, achieving a high AUC (0.8979) and a significantly better AUCPR (0.7683). This indicates that Random Forest can effectively handle the imbalanced dataset by using an ensemble of decision trees to capture complex patterns. However, while it outperforms simpler models, it still falls short of the top-performing XGBoost model.

5.6 XGBoost

- AUC: 0.9770
- AUCPR: 0.8618

  XGBoost outperforms all other models, achieving the highest AUC (0.9770) and AUCPR (0.8618). This demonstrates its strong capability in identifying fraudulent transactions with high precision and recall. XGBoost's use of gradient boosting, regularization, and optimization techniques allows it to handle the class imbalance effectively while maintaining high predictive power.

Key Takeaways:

1. **Best Model**: XGBoost is the best-performing model, with both the highest AUC and AUCPR scores, indicating its robustness in distinguishing fraudulent transactions even in a highly imbalanced setting.

2. **Random Forest**: While not as powerful as XGBoost, the Random Forest model provides a good balance between performance and interpretability, making it a viable option for practical implementations.

3. **Poor Performers**: Naive Bayes, KNN, and SVM show limitations in detecting fraud due to their assumptions, sensitivity to data distributions, or inability to handle the extreme imbalance effectively.

4. **Importance of AUCPR**: In the context of fraud detection, AUCPR is a more critical metric than AUC, as it focuses on the model's ability to detect the minority class (fraud) accurately. This is why models like XGBoost, with a high AUCPR, are preferred for fraud detection tasks.

6. Conclusion

The results highlight the superiority of ensemble-based models like XGBoost and Random Forest in handling imbalanced datasets. These models leverage their ability to capture complex interactions among features and optimize for both precision and recall, making them the most suitable choices for real-world fraud detection systems.

This report demonstrated the effectiveness of machine learning models in identifying fraudulent transactions in an imbalanced dataset. The XGBoost model outperformed Random Forest, achieving an AUCPR of 0.8618, which indicates a strong ability to identify fraud while minimizing false positives.

6.1 Potential Impact

Implementing such models in real-time systems can significantly reduce financial losses due to fraud. However, the models must be continuously monitored and retrained to adapt to evolving fraud patterns.

6.2 Limitations

- The dataset is anonymized, limiting the ability to leverage domain-specific features like user demographics.
- Synthetic oversampling techniques like SMOTE may introduce bias if not carefully monitored.

6.3 Future Work

1. **Advanced Techniques**: Explore Conditional Generative Adversarial Networks (GANs) for generating more realistic synthetic fraud samples.
2. **Real-Time Deployment**: Deploy models on platforms like AWS SageMaker for real-time fraud detection [15].

References

1. Kaggle - Credit Card Fraud Detection Dataset. Retrieved from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

2. scikit-learn Documentation - Supervised Learning Algorithms. Retrieved from https://scikit-learn.org/stable/supervised_learning.html

3. XGBoost Documentation - A Scalable Tree Boosting System. Retrieved from https://xgboost.readthedocs.io/en/stable/

4. A Method to Perform Synthetic Minority Over-sampling Technique (SMOTE). Retrieved from https://imbalanced-learn.org/stable/over_sampling.html

5. Random Forests for Classification and Regression. Retrieved from https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

6. Understanding Support Vector Machines (SVM) for Classification. Retrieved from https://scikit-learn.org/stable/modules/svm.html

7. An Overview of Naive Bayes Algorithm. Retrieved from https://scikit-learn.org/stable/modules/naive_bayes.html

8. Understanding Precision-Recall Curves. Retrieved from https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

9. Gradient Boosting for Classification. Retrieved from https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting

10. Best Practices for Data Cleaning in Machine Learning. Retrieved from https://www.tableau.com/learn/articles/what-is-data-cleaning

11. Techniques for Feature Engineering in Machine Learning. Retrieved from https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

12. Python Data Analysis with Pandas Documentation. Retrieved from https://pandas.pydata.org/docs/

13. Data Visualization with Matplotlib: Creating Visuals. Retrieved from https://matplotlib.org/stable/contents.html

14. Evaluating Machine Learning Models: A Beginner's Guide. Retrieved from https://medium.com/@sachin.rawat_85554/machine-learning-development-a-beginners-guide-to-development-f48b7932cf13

15. Advances in Fraud Detection Using Machine Learning. Retrieved from https://www.datrics.ai/articles/machine-learning-for-fraud-detection