Credit Card Fraud Detection

Rania Anwar
Supervisor Name: Tamer Abdou, PhD
Submission Date: July 24, 2023

Ryerson University

Table of Contents

01

Understanding Credit Card Fraud

Define credit card fraud (CCF), briefly visualize credit card transactions data, analyze existing literature on the topic

02

Objective and Research Questions

Outline the objectives of the project followed by proposing the research questions that will be investigated

03

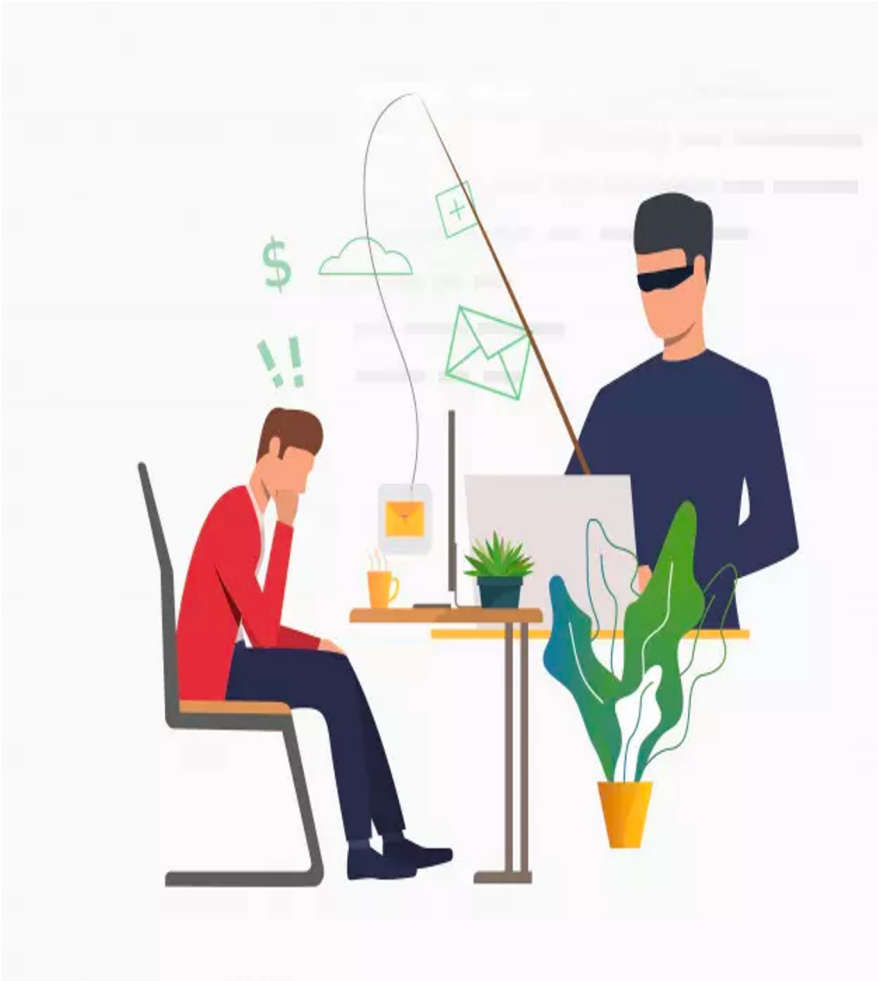
Methodology

Discuss the approach taken to answer research questions along with the tools necessary

04

Results

Elaborate on the key findings, compare results from the literature, highlight limitations and recommendations



Source: Sakasegawa (2022)

Ryerson University

01

Understanding Credit Card Fraud

Ryerson University

What is Credit Card Fraud (CCF)

- Credit card fraud occurs when a person uses a stolen credit or information to make unauthorized purchases under the person's credit card name (Rafter, 2017)
- Several approaches for fraudsters to attain one's credit card information
  - Fake websites (Rafter, 2017)
  - Phishing emails (Dickey, 2021)
  - Data breaches (Rafter, 2017)

The State of Credit Card Fraud

- "Three-in-five 18-34 year olds (63%) report being a victim of at least one type of financial fraud in their lifetime" (Chartered Professional Accountants of Canada, 2023)
- 72% of participants in the study with credit cards state that they manage their finances online (Chartered Professional Accountants of Canada, 2023)
- "Costs the global economy $5.127 trillion annually" (Falco, 2023)
  - 72% of businesses globally reported an increase in fraud within 12 months

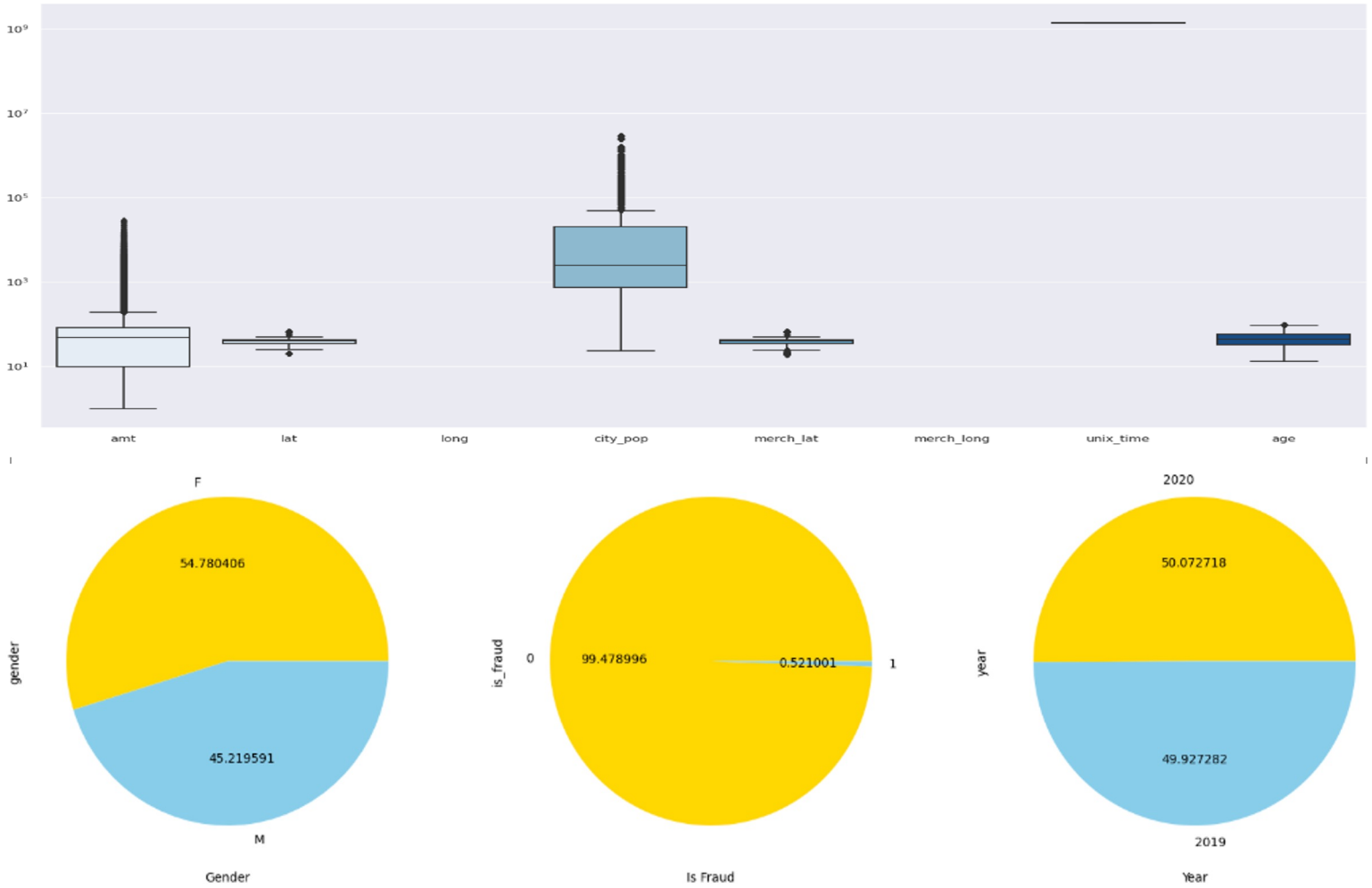
Ryerson University

Data Science and Credit Card Fraud

- Data science is significantly involved in collecting, summarizing, and predicting banking data to identify suspicious activities (Qualetics Team, 2019)
  - Models are created to classify credit card transactions based on features such as amount, location, merchant, month, etc.
- Machine learning plays a critical role in fraud detection (Kosourova, 2022)
  - Approached through supervised or unsupervised methods

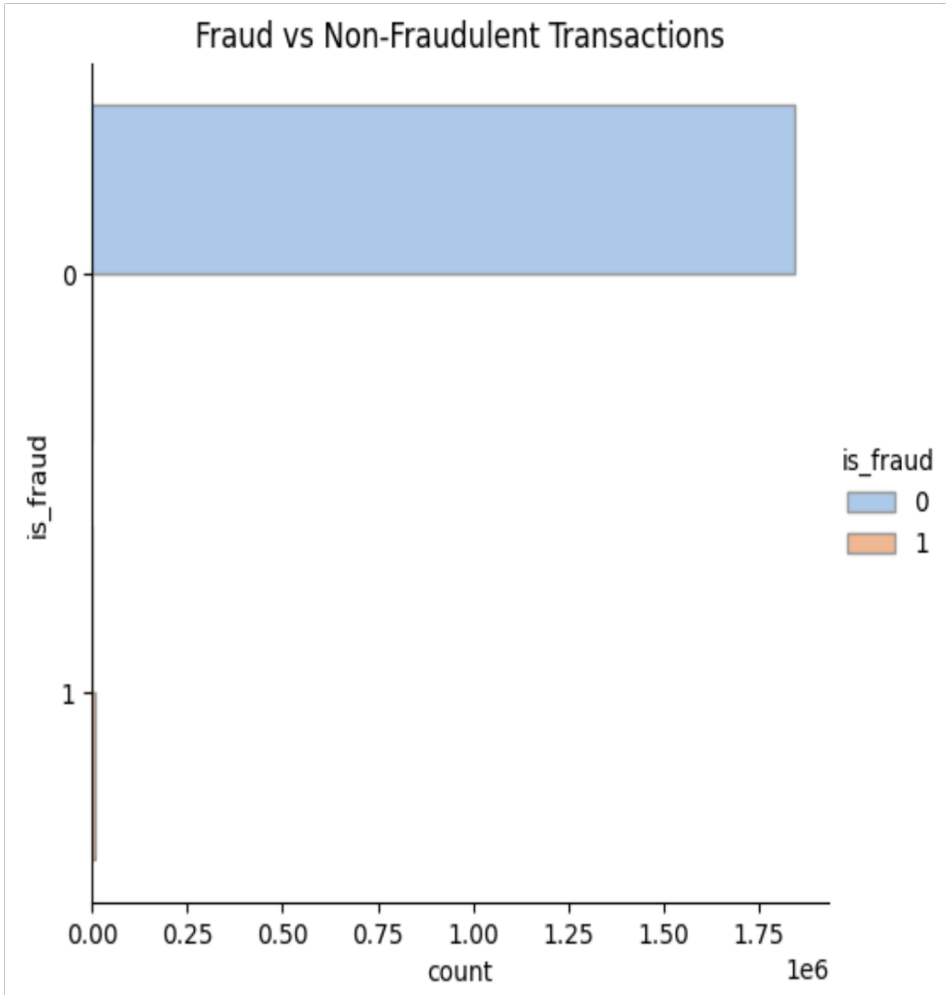
Ryerson University

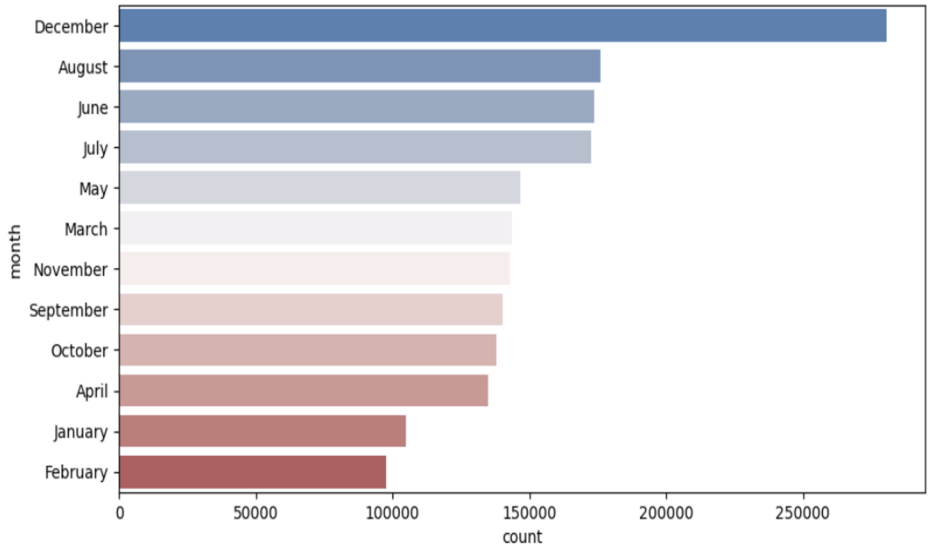
Exploring Credit Card Transaction Data

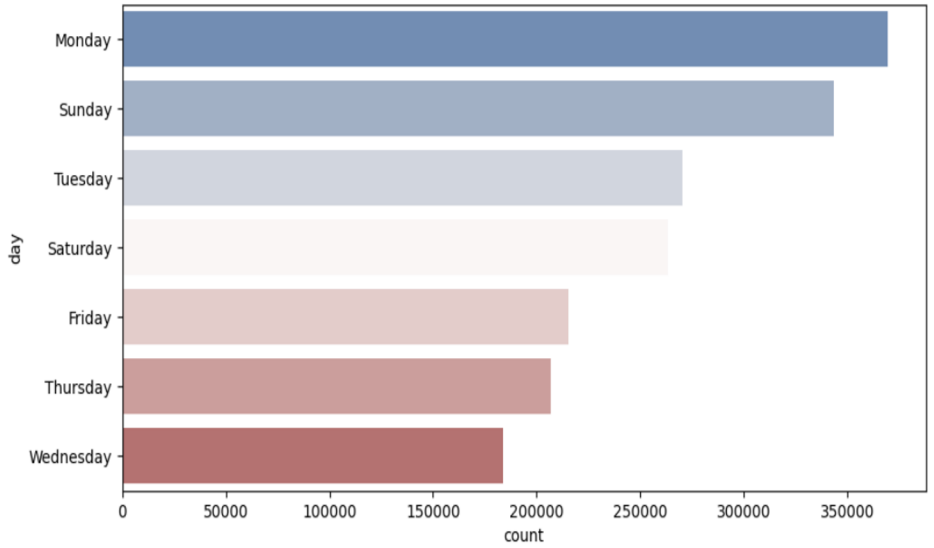


Ryerson University

Exploring Credit Card Transaction Data

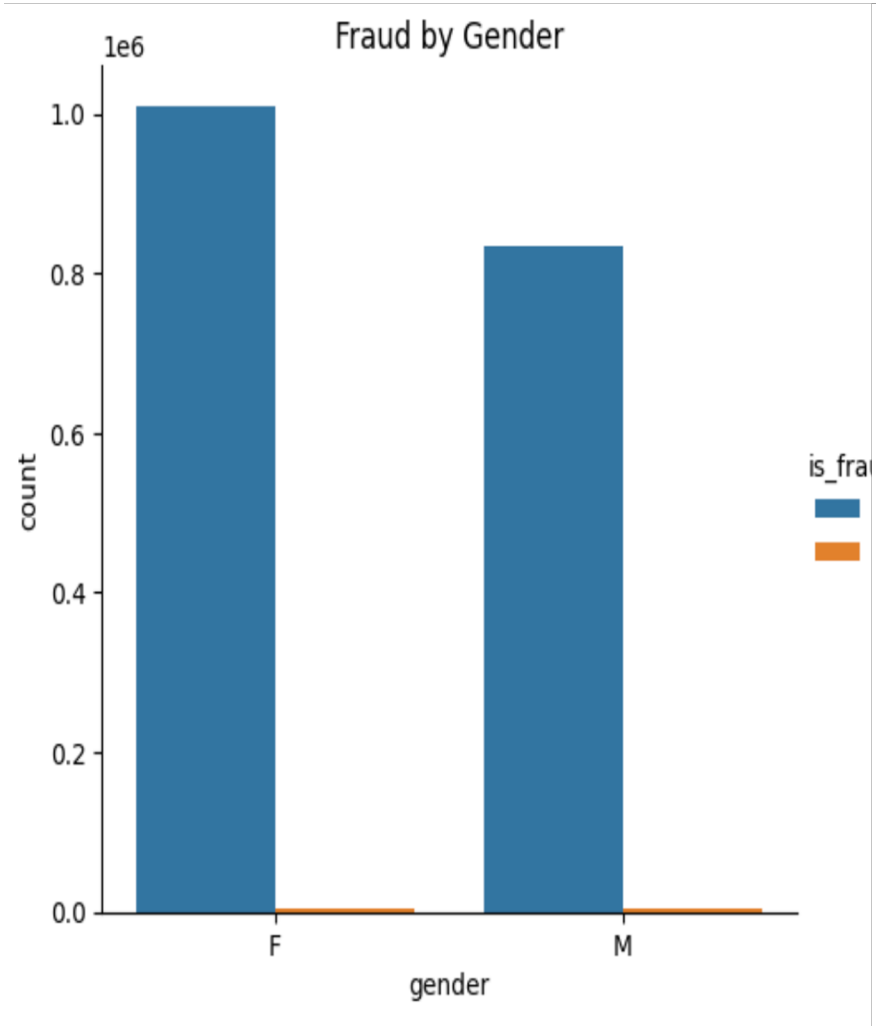


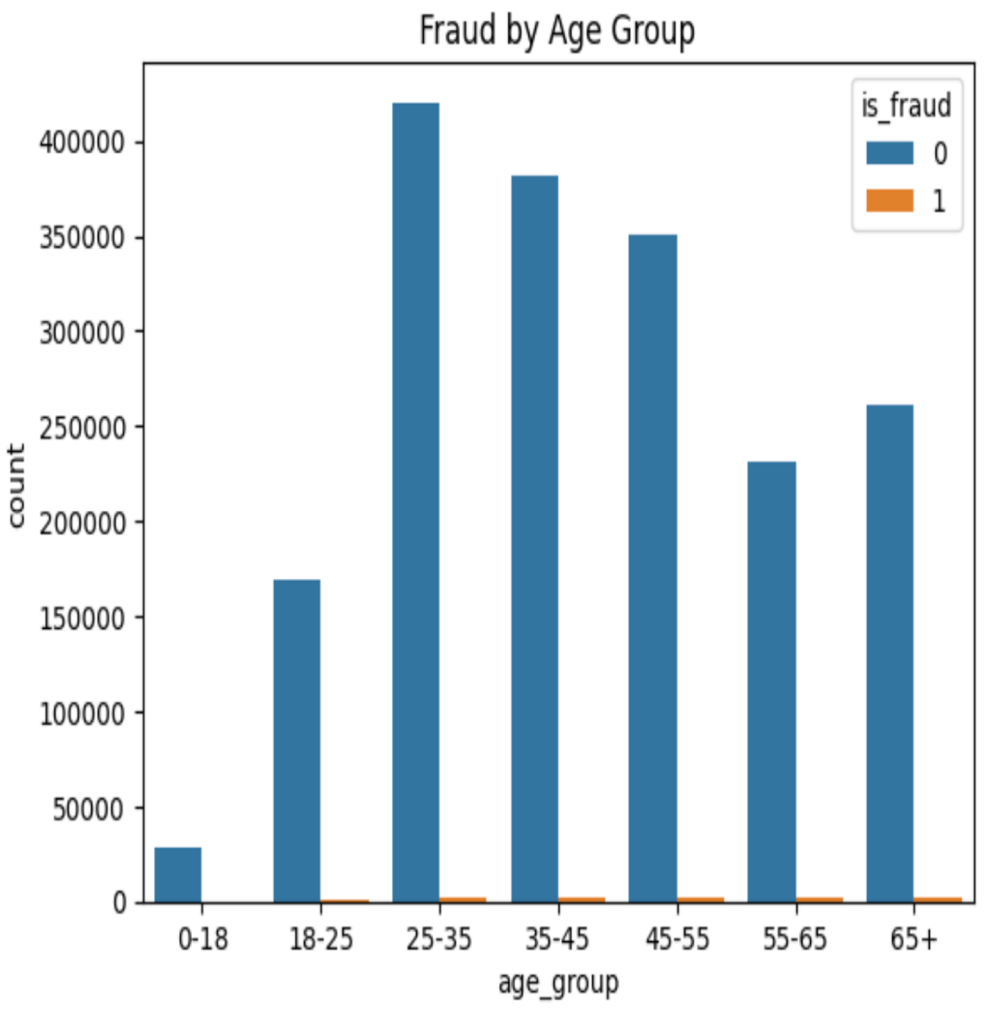




Ryerson University

Exploring Credit Card Transaction Data





Ryerson University

Literature Review

- Using SMOTE to combat class imbalance, Haddab (2022) employed logistic regression, random forest, Naive Bayes, and multilayer perceptron algorithms to select the best performing classifier
  - Random forest had the best performance with an accuracy of 98.96%, precision of 97.38%, and recall of 83.63%
- Also with the goal to find the best performing classifier, Afriyie et al. (2023) used logistic regression, decision tree, and random forest along with undersampling the data
  - Random forest performed best with an AUC of 98.8%

Literature Review

- Alabrah (2023) compares five classifiers with and without the normalization of outliers to detect differences in performance
  - Machine learning models in the second experiment with outliers normalized provided more robust results
- Mustaqim et al. (2021) studied the effectiveness of feature selection and its influence on classification results by employing recursive feature elimination with cross validation (RFECV)
  - Found that applying different values of k lead to lower accuracy in decision tree model but no effect in Naive Bayes model
  - Increase in k resulted in lower number of selected features

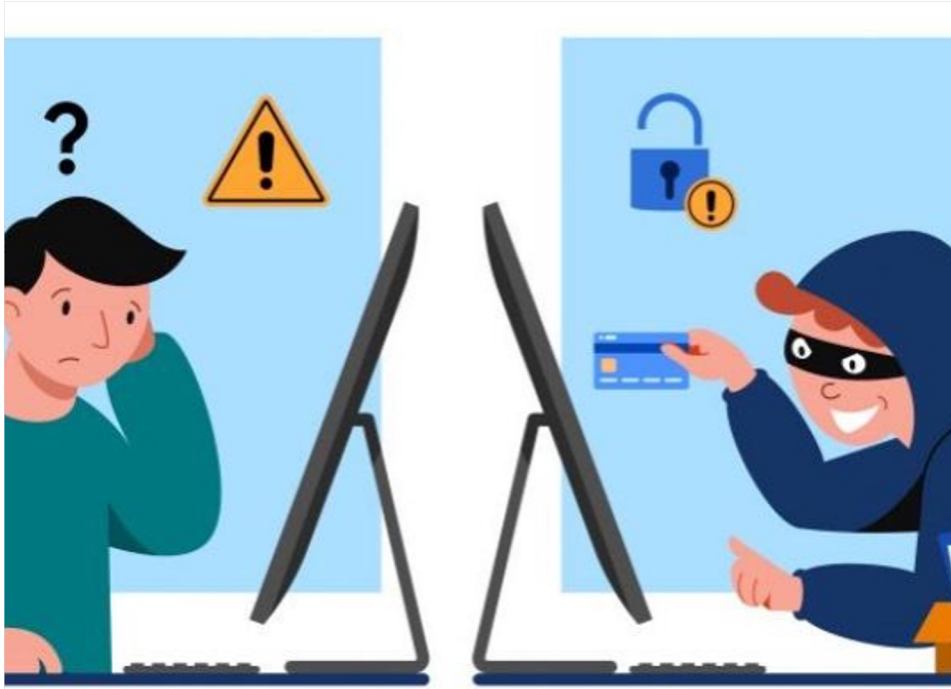
02

Objectives and Research Questions

Ryerson University

Objectives

- Compare three resampling techniques with three machine learning algorithms to identify the best performing model
- Examine differences in performance of models with and without outliers
- Observe performances of models with all and selected features



Source: Gareth (2022)

Ryerson University

Research Questions

1. Knowing that there exists a bias towards non-fraudulent transactions, what sampling technique will perform better for creating an accurate fraud detection model?
2. In relation to the first question, will the models perform better when outliers are present and transformed, or when no outliers exist?
3. What are the essential features that assist in determining the two classes of transactions, legitimate and fraudulent?

Ryerson University

03

Methodology

Ryerson University

Data Description

- "A simulated credit card transaction dataset containing legitimate and fraud transactions from the duration of January 1st to December 31st, 2020" (Shenoy, 2019)
  - Generated using Sparkov Data Generation tool designed by Brandon Harris
  - Contains credit cards of 1000 consumers in the United States and 800 merchants
  - 1,296,675 observations with 23 features

Ryerson University

Dataset

	cc_num	merchant	category	amt	first	last	gender	street	city	state
0	0 2703186189652095\n1 6304...	fraud_Rippin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	Moravian Falls	NC
1	0 2703186189652095\n1 6304...	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	Orient	WA
2	0 2703186189652095\n1 6304...	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	Malad City	ID
3	0 2703186189652095\n1 6304...	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	Boulder	MT
4	0 2703186189652095\n1 6304...	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	Doe Hill	VA

Ryerson University

Methodology

- **Data Preparation**
  - Reading in the datasets into pandas DataFrame, data type conversions, feature engineering, dropping redundant variables
- **Exploratory Data Analysis (EDA)**
  - Univariate and bivariate analysis, data visualization (bar plots, box plots, pie charts)
  - Revisit data preparation stage
- **Data Preprocessing**
  - Create copy of dataset
  - Scaling the datasets (RobustScaler on original dataset and MinMaxScaler on copied dataset)
  - One-hot encoding of categorical features, train-test split (70% training, 30% testing), feature selection using RFECV

Methodology

- **Data Modelling**
  - Pipelines to chain resampling technique with machine learning algorithm
    - Resampling techniques: random under sampling, random over sampling, synthetic minority over sampling (SMOTE)
    - Machine learning algorithms: logistic regression, decision tree, random forest
  - Hyperparameter tuning with RandomizedSearchcv
    - Fitted with the training set
- **Performance**
  - Performance metrics: accuracy, recall, precision, Matthew's correlation (MCC), F1 score
  - Receiver operating characteristic curves (ROC) and area under the curve value (AUC)

Tools

- **Pandas**
  - Open source package for data analysis and manipulation (Pandas, 2023)
- **Scikit-learn (Sklearn)**
  - Open source package for machine learning (Pedregosa et al., 2011)
- **imblearn**
  - Open source MIT-licensed library for managing imbalanced data (The imbalance-learn developers, 2023)
- **NumPy**
  - Scientific computing package to perform operations on arrays (NumPy Developers, 2022)
- **Matplotlib**
  - Data visualization package for creating plots (The Matplotlib development team, 2023)
- **Seaborn**
  - High level interface data visualization package for graphing (Waskom, 2022)

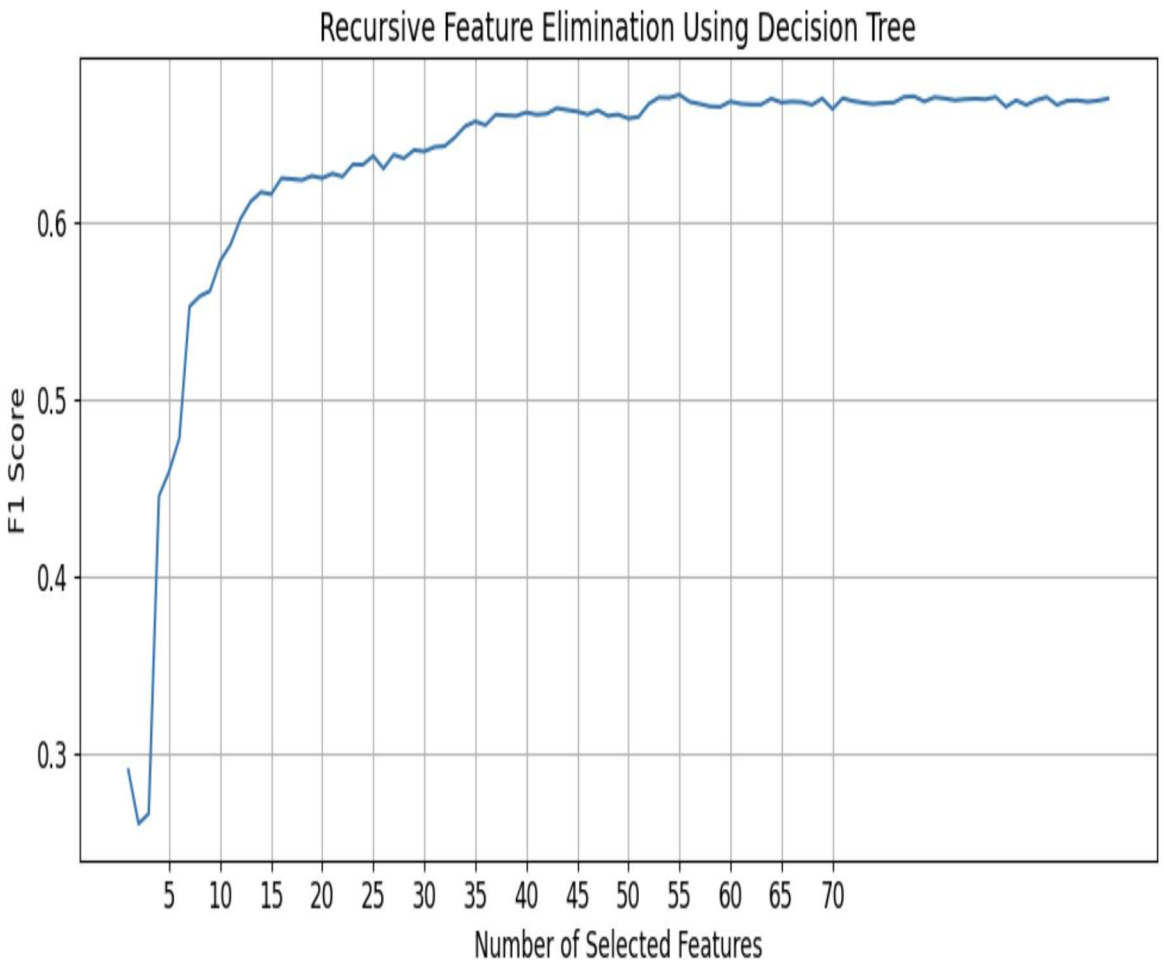
04

Results

Ryerson University

Feature Selection

- Best performing number of features was 55
- Features selected included:
  - amt (amount)
  - gender
  - lat (latitude)
  - long (longitude)
  - year
  - age



Ryerson University

Performance of models on original dataset using RobustScalar containing all features vs selected features

Table 1: Performance of using random under sampling on three machine learning algorithms on original dataset.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.96	0.75	0.96	0.09	0.99	0.25	0.16	0.86
Decision Tree	0.96	0.96	0.96	0.11	0.99	0.32	0.20	0.96
Random Forest	0.98	0.93	0.98	0.21	0.99	0.44	0.34	0.96

Table 2: Performance of using random over sampling on three machine learning algorithms on original dataset.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.90	0.75	0.90	0.04	0.99	0.16	0.07	0.83
Decision Tree	0.99	0.79	0.99	0.32	0.99	0.50	0.46	0.89
Random Forest	0.99	0.82	0.99	0.67	0.99	0.74	0.74	0.91

Table 3: Performance of using SMOTE on three machine learning algorithms on original dataset.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.98	0.69	0.99	0.21	0.99	0.37	0.31	0.84
Decision Tree	0.99	0.87	0.99	0.31	0.99	0.52	0.46	0.93
Random Forest	0.99	0.83	0.99	0.55	0.99	0.71	0.71	0.91

Table 7: Performance of using random under sampling on three machine learning algorithms on original dataset with selected features.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.96	0.75	0.96	0.09	0.99	0.25	0.16	0.86
Decision Tree	0.95	0.95	0.95	0.09	0.99	0.28	0.16	0.95
Random Forest	0.98	0.92	0.98	0.23	0.99	0.46	0.37	0.95

Table 8: Performance of using random over sampling on three machine learning algorithms on original dataset with selected features.

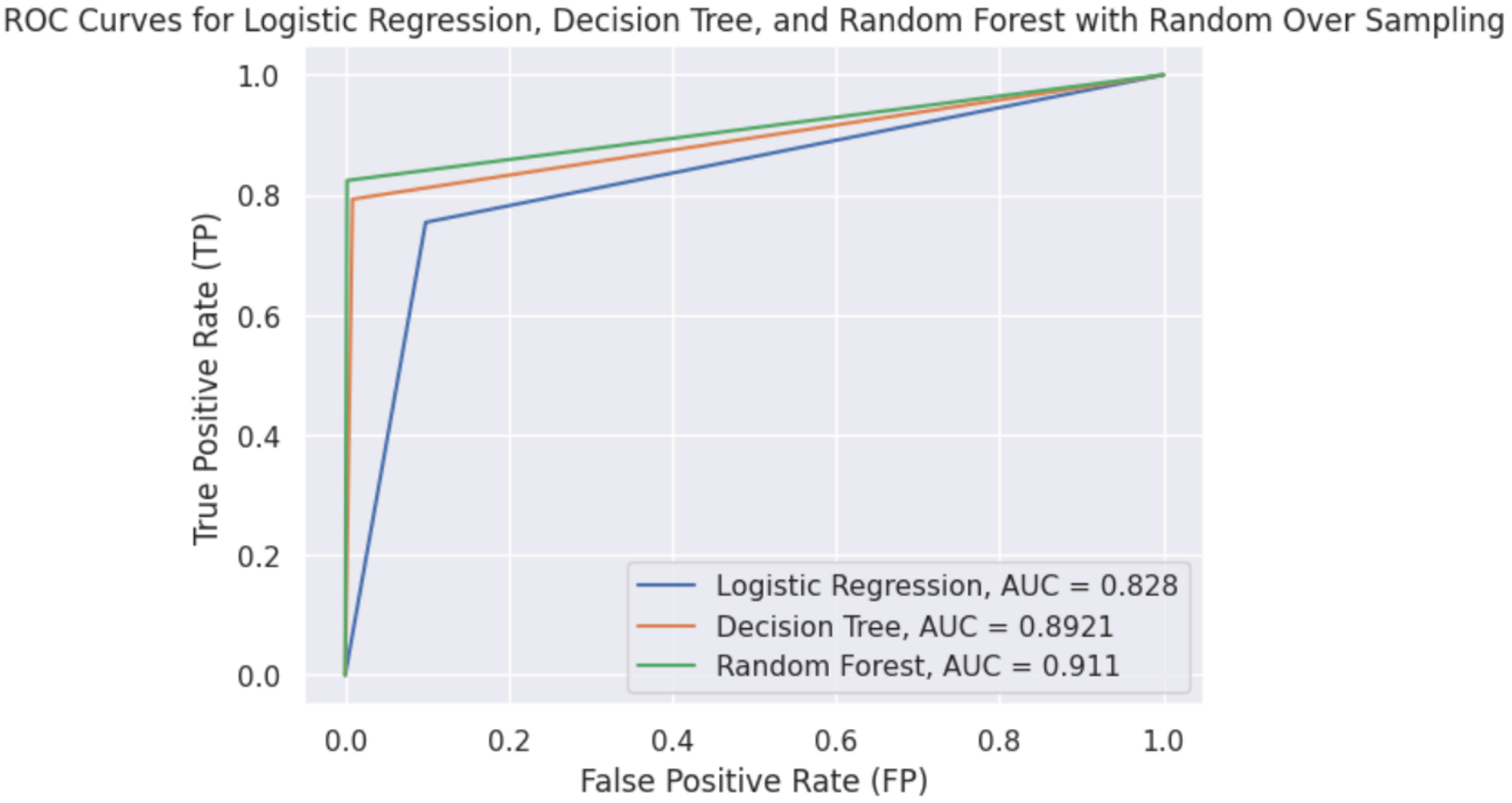
	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.90	0.76	0.90	0.04	0.99	0.16	0.08	0.83
Decision Tree	0.98	0.85	0.98	0.23	0.99	0.43	0.36	0.92
Random Forest	0.99	0.86	0.99	0.56	0.99	0.69	0.68	0.93

Table 9: Performance of using SMOTE on three machine learning algorithms on original dataset with selected features.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.98	0.69	0.98	0.14	0.99	0.31	0.23	0.83
Decision Tree	0.97	0.80	0.97	0.12	0.99	0.30	0.21	0.88
Random Forest	0.99	0.81	0.99	0.38	0.99	0.56	0.52	0.91

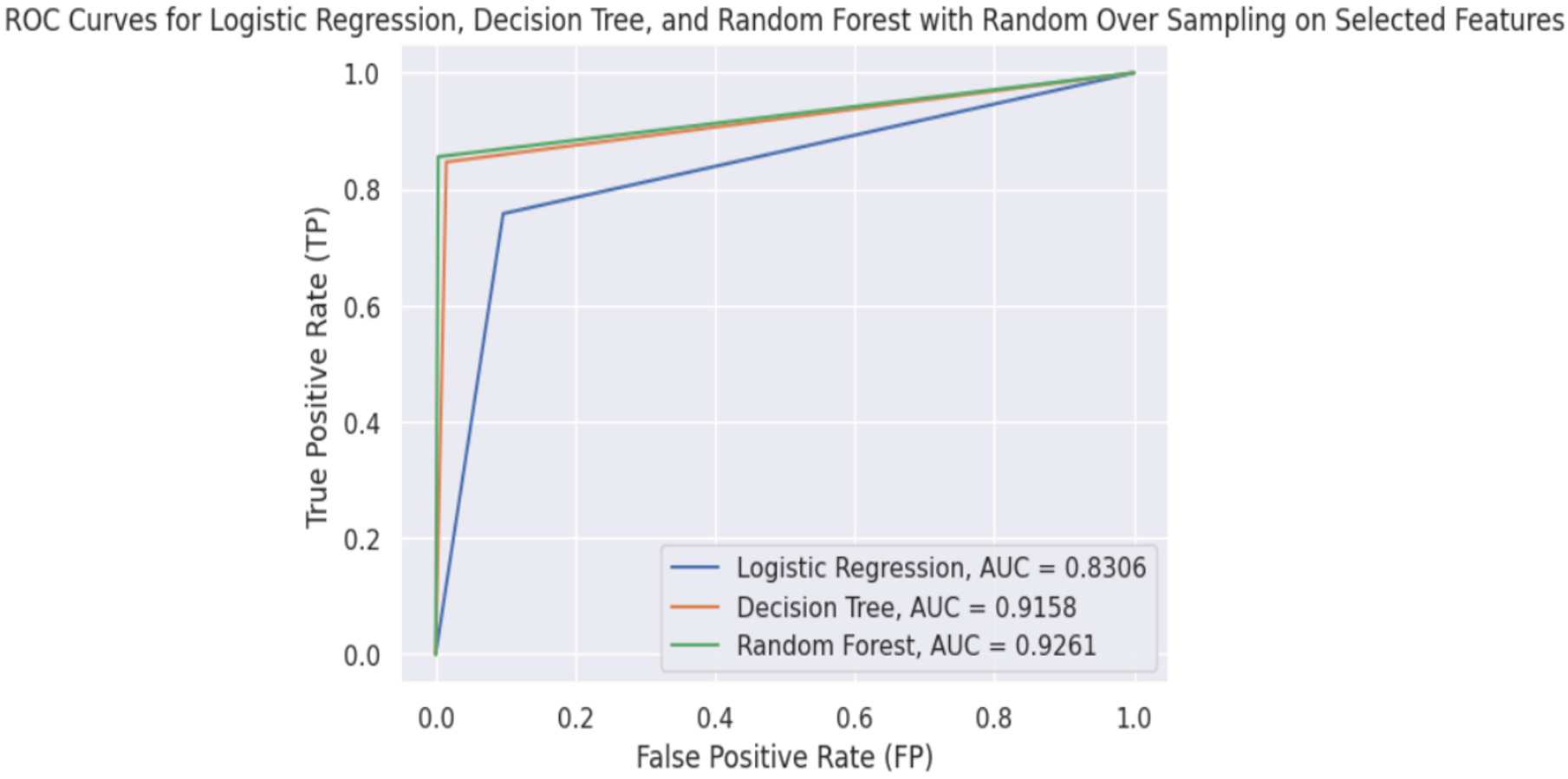
Ryerson University

ROC Curves for Random Over Sampling



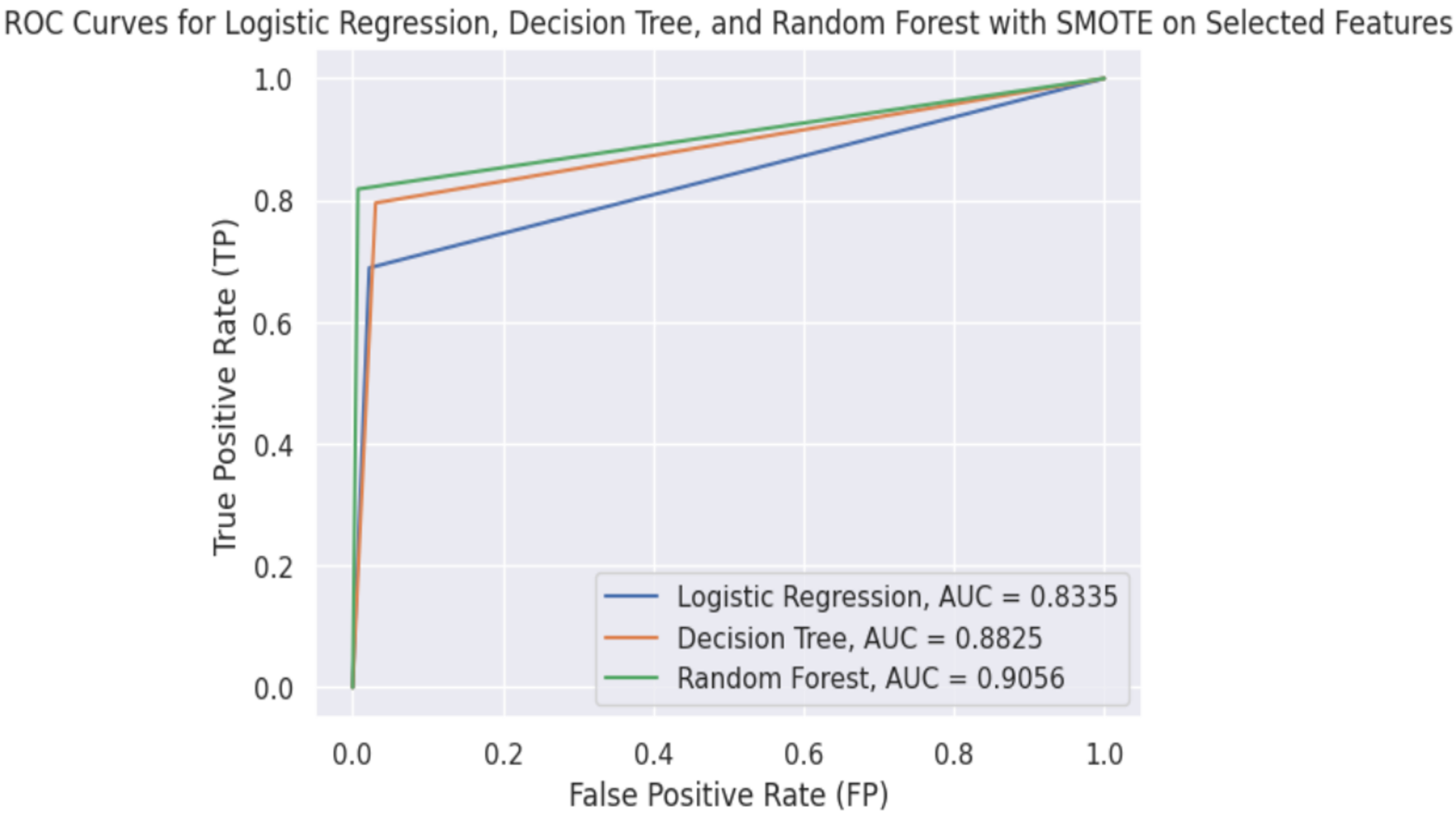
Ryerson University

ROC Curves for Random Over Sampling (Selected Features)



Ryerson University

ROC Curves for SMOTE (Selected Features)



Ryerson University

Performance of models with no outliers containing all features vs selected features

Table 4: Performance of using random under sampling on three machine learning algorithms on copied dataset containing no outliers.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.71	0.72	0.71	0.00	0.99	0.04	0.00	0.72
Decision Tree	0.92	0.96	0.92	0.02	0.99	0.12	0.03	0.94
Random Forest	0.89	0.97	0.89	0.01	0.99	0.10	0.03	0.93

Table 5: Performance of using random over sampling on three machine learning algorithms on copied dataset containing no outliers.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.66	0.89	0.66	0.00	0.99	0.04	0.00	0.77
Decision Tree	0.99	0.65	0.99	0.08	0.99	0.23	0.15	0.82
Random Forest	0.99	0.62	0.99	0.14	0.99	0.30	0.23	0.81

Table 6: Performance of using SMOTE on three machine learning algorithms on copied dataset containing no outliers.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.97	0.27	0.97	0.01	0.99	0.05	0.02	0.62
Decision Tree	0.99	0.62	0.99	0.07	0.99	0.20	0.12	0.80
Random Forest	0.98	0.70	0.98	0.05	0.99	0.18	0.09	0.84

Table 10: Performance of using random under sampling on three machine learning algorithms on copied dataset containing no outliers with selected features.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.73	0.69	0.73	0.00	0.99	0.04	0.00	0.71
Decision Tree	0.89	0.91	0.89	0.01	0.99	0.09	0.02	0.90
Random Forest	0.89	0.94	0.89	0.01	0.99	0.10	0.02	0.92

Table 11: Performance of using random over sampling on three machine learning algorithms on copied dataset containing no outliers with selected features.

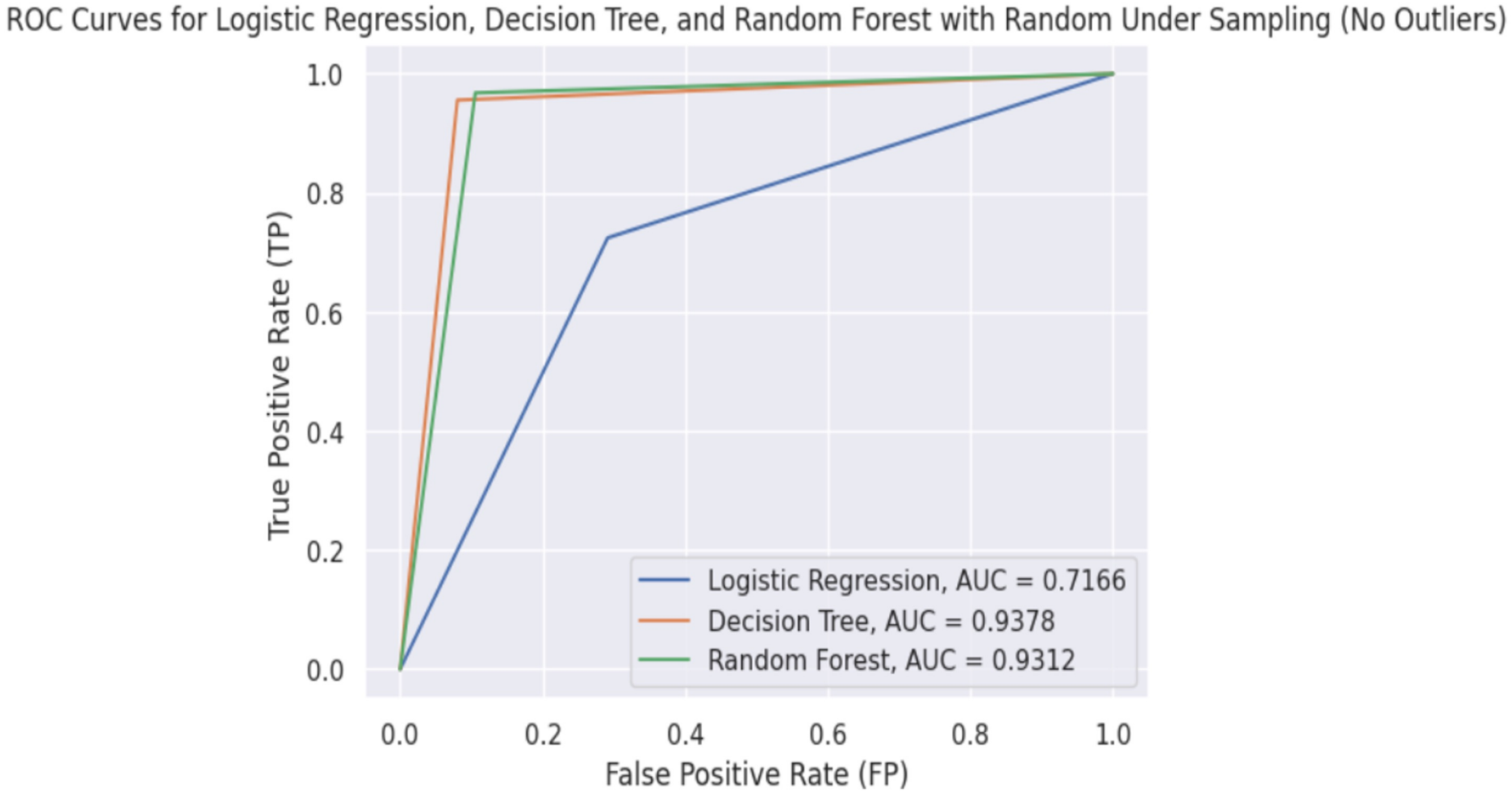
	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.62	0.90	0.62	0.00	0.99	0.04	0.00	0.76
Decision Tree	0.99	0.59	0.99	0.05	0.99	0.18	0.10	0.79
Random Forest	0.99	0.60	0.99	0.11	0.99	0.26	0.19	0.80

Table 12: Performance of using SMOTE on three machine learning algorithms on copied dataset containing no outliers with selected features.

	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)	MCC	F1 Score	AUC
Logistic Regression	0.81	0.59	0.81	0.00	0.99	0.04	0.00	0.70
Decision Tree	0.94	0.72	0.94	0.02	0.99	0.10	0.03	0.83
Random Forest	0.93	0.75	0.93	0.02	0.99	0.10	0.03	0.84

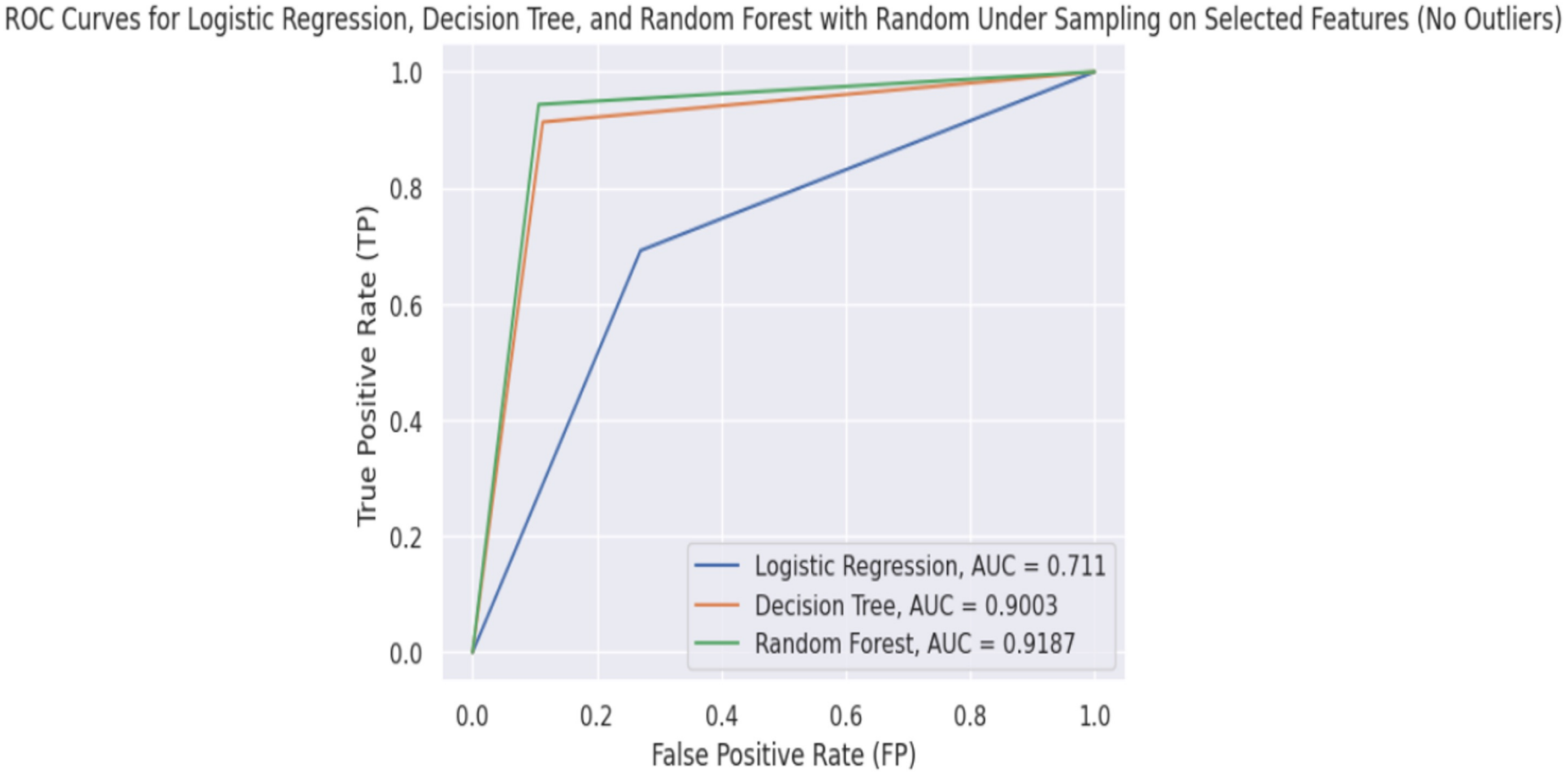
Ryerson University

ROC Curves for Random Under sampling



Ryerson University

ROC Curves for Random Under sampling (Selected Features)



Ryerson University

Comparing Results with Previous Studies

- In this study, models created with original dataset that underwent under sampling and containing all features (Table 1) achieved somewhat better performance than Afriyie et al. (2023)
  - Logistic regression and decision tree model had an accuracy of approximately 96% in comparison to 92% in their study
  - F1 score of models in both studies were however relatively weak
  - Recall of the positive class were relatively similar

Table 11
Comparing the models' performances.

Model name	Accuracy	F1-Score	Recall	Precision	Specificity
Decision tree	0.92	0.09	0.93	0.05	0.92
Random forest	0.96	0.17	0.97	0.09	0.96
Logistics regression	0.92	0.08	0.76	0.04	0.92

Source: Afriye et al. (2023)

Comparing Results with Previous Studies

- Similar to what was observed in Alabrah's (2023) study, models created with the original dataset using RobustScalar had higher performance in comparison to models containing no outliers in terms of precision, MCC, and F1 score
- Almost in line with the findings of Afriyie et al. (2023) and Haddab (2022), random forest with random over sampling was the best performing classifier
  - With criterion set to entropy and a maximum depth of 25

Ryerson University

Limitations and Recommendations

- Selection of hyperparameter tuning algorithm
  - Only values selected by the user enter the grid space, leaving out other potential values that offer a more optimal set
  - Adopt Bayesian optimization as a potential alternative (Gupta, 2020)
- Utilizing ROC AUC as a performance metric
  - Provided overly optimistic results, biased towards the negative class, which is legitimate transactions
  - Employ instead precision-recall curves since more emphasis is given to the minority class (Brownlee, 2020)
- Examined limited machine learning algorithms
  - Other supervised machine learning algorithms may be better classifiers for distinguishing the two types of transactions
  - Consider Support Vector Machine (SVM), Naive Bayes, and among others

Conclusion

- Creating a fraud detection model is a continuous process
- Explored three resampling techniques with three machine learning algorithms, examined performance of models with and without outliers and with selected features
  - Three resampling techniques employed were random under sampling, random over sampling, and SMOTE
  - Three machine learning algorithms were logistic regression, decision tree, and random forest
  - Applied Sklearn RobustScalar on original dataset and MinMaxScalar for copied dataset
  - RFECV for feature selection

Ryerson University

Conclusion

- Models that were created using the original dataset with RobustScalar and with all features overall achieved better performance across all metrics in comparison to the copied dataset with no outliers
- Specifically, over sampling the data and using random forest on the original dataset achieved the highest performance

Ryerson University

References

Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., ... & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163.

Alabrah, A. (2023). An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method. Sensors, 23(9), 4406.

Brownlee, J. (2020, January 5). ROC Curves and Precision-Recall Curves for Imbalanced Classification - MachineLearningMastery.com. MachineLearningMastery.com. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Chartered Professional Accountants of Canada. (2023, February 15). Unlikely targets: More young Canadians report being a victim of financial fraud than older Canadians: CPA Canada survey reveals. Cpacanada.ca. https://www.cpacanada.ca/en/the-cpa-profession/about-cpa-canada/media-centre/2023/february/cpa-canada-fraud-survey-2023

Dickey, C. (2023, March 13). Ways your credit card info might be stolen and how to prevent it. Bankrate; Bankrate.com. https://www.bankrate.com/finance/credit-cards/5-ways-theives-steal-credit-card-data/

Ryerson University

References

Falco, A. (2023). Understanding the Threat of Card Transaction Fraud and its Impact on the Financial Ecosystem | Waylay Blog. Waylay.io. https://www.waylay.io/articles/understanding-the-threat-of-card-transaction-fraud-and-its-impact-on-the-financial-ecosystem#:~:text=Globally%2C%20card%20fraud%20is%20making,less%20than%2024%20hours%20old.

Gupta, L. (2020, November 21). Comparison of Hyperparameter Tuning algorithms: Grid search, Random search, Bayesian optimization. Medium; Analytics Vidhya. https://medium.com/analytics-vidhya/comparison-of-hyperparameter-tuning-algorithms-grid-search-random-search-bayesian-optimization-5326aaef1bd1

Haddab, D. M. (2022). Data Science & Machine Learning Methods for Detecting Credit Card Fraud. International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), 4(4), 71-75.

Kosourova, E. (2022, March 21). Data Science in Banking: Fraud Detection. Datacamp.com; DataCamp. https://www.datacamp.com/blog/data-science-in-banking

Mustaqim, A. Z., Adi, S., Pristyanto, Y., & Astuti, Y. (2021, June). The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. In 2021 International conference on artificial intelligence and computer science technology (ICAICST) (pp. 270-275). IEEE.

Ryerson University

References

NumPy Developers. (2022). What is NumPy? — NumPy v1.25 Manual. Numpy.org. https://numpy.org/doc/stable/user/whatisnumpy.html

Pandas. (2023). pandas - Python Data Analysis Library. Pydata.org. https://pandas.pydata.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Qualetics Team. (2019, October 28). *Data Science In Banking: 5 Use Cases For Banks*. Qualetics Data Machines| Data Analytics, AI | Enable Analytics & AI for Web and Mobile Applications. https://qualetics.com/data-science-in-banking-5-use-cases-for-banks/

Rafter, D. (2017, September 14). What Is Credit Card Fraud? LifeLock by Norton. https://lifelock.norton.com/learn/fraud/what-is-credit-card-fraud

Shenoy, K. (2019). Credit Card Transactions Fraud Detection Dataset. Kaggle.com. https://www.kaggle.com/datasets/kartik2112/fraud-detection

The imbalanced-learn developers. (2023). imbalanced-learn documentation — Version 0.11.0. Imbalanced-Learn.org. https://imbalanced-learn.org/stable/

Ryerson University

References

The Matplotlib development team . (2023). Matplotlib — Visualization with Python. Matplotlib.org. https://matplotlib.org/

Waskom, M. (2022). seaborn: statistical data visualization — seaborn 0.12.2 documentation. Pydata.org. https://seaborn.pydata.org/

Ryerson University