040172-1 (2024WS)
VU Doing Data Science

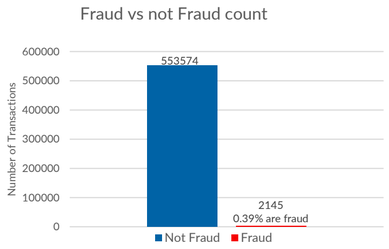
Group 5: Credit Card Fraud Prediction

Elizaveta Kotikova, Martyna Janina Kopyta, Péter Ferenc Gyarmati

Dataset

Customer and Transaction Data

- 555,719 instances
- 22 attributes
- Mix of categorical and numerical data types
- No null values
- **Imbalanced** dataset



• Timestamp of the transaction • Customer id number • Merchant • Category Transaction type • Transaction amount • Cardholder's first and last name• Cardholder's address• Latitude and Longitude of cardholder's location • Population of the cardholder's city • Cardholder's date of birth• Unique transaction id• Transaction timestamp…

You can find the **documentation of the entire process** here:

Challenge

Why is it challenging?

- High cost of errors
- Fraudsters evolution
- Real-Time detection required
- Personal data → privacy regulations

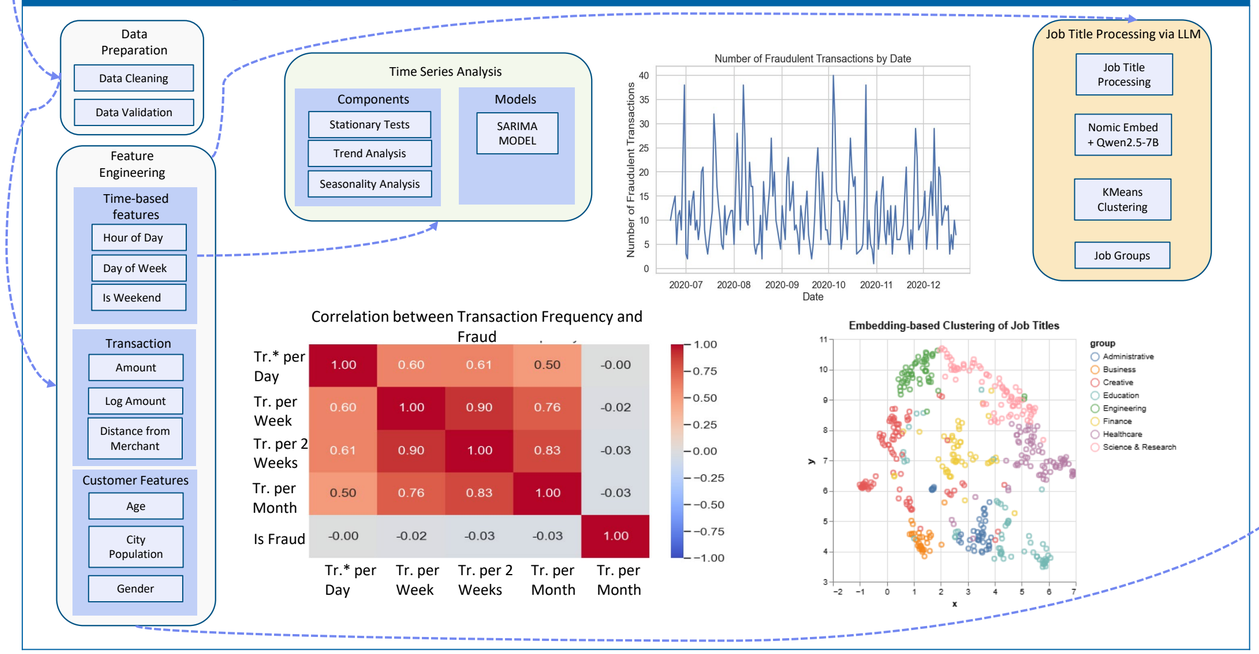
What simplifies the analysis?

- No domain knowledge required

Expected benefits :

- Reduced financial losses
- Increased trust and satisfaction
- Efficient fraud investigation

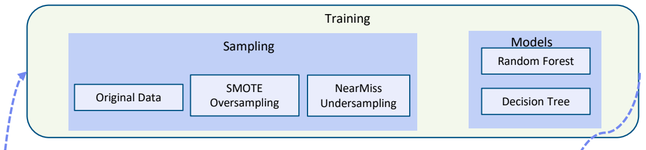
Data Understanding & Preparation



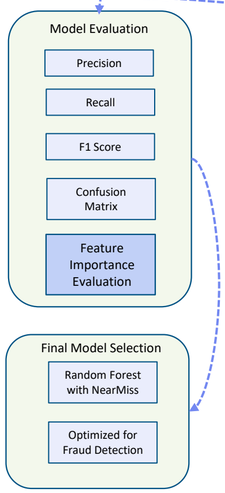
Modeling

Classification

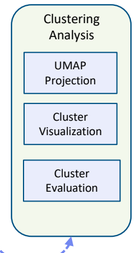
- 3 datasets: original, SMOTE-oversampled, and NearMiss-undersampled.
- Training of decision tree and random forest classifiers.
- → 6 models with varying Class-wise Performance.



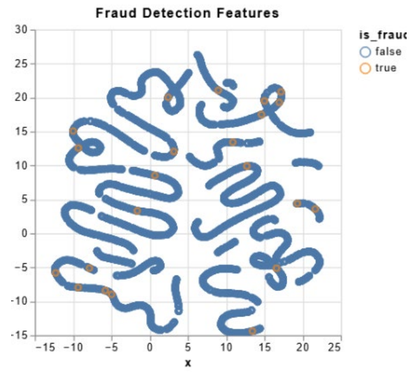
Model	Overall Metrics	Feature Importance
Decision Tree, Original Samples	High accuracy, imbalance issues evident	Transaction amount, category, time
Random Forest, Original Samples	High macro precision, lower macro recall	Transaction amount, time
Random Forest, SMOTE Oversampling	Good macro precision, weaker recall, trade-off	Transaction amount, gender, merchant frequency
Random Forest, NearMiss Undersampling	Better recall and accuracy, limited precision	Transaction amount, time-related factors



Clustering



- Visualization of prepared features in two-dimensional space using UMAP.
- Exploration with KMeans, DBSCAN, and agglomerative hierarchical clustering.
- None were effective in distinguishing fraudulent transactions from legitimate ones.



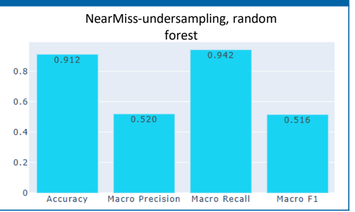
Results

Random Forests: higher precision.

Decision Trees: simpler, faster, task-dependent.

SMOTE: improves recall by reducing false negatives.

NearMiss: balances errors, higher fraud detection with more false positives.



Conclusion

Clustering

Explored KMeans, DBSCAN, hierarchical clustering—none effectively distinguished fraud.

→Shifted to supervised learning.

Classification

Random Forest or Decision Tree trained on the undersampled dataset with NearMiss might be the best choice.

University of Vienna, Department of Business Analytics