Bank Account Fraud

Name: Kai Lu
Toronto Met ID: 501217886
Supervisor: Dr. Tamer Abdou

Ryerson University

Project Objectives

- Can we use **traditional machine learning** (ML) models to detect online bank account opening fraud?

- After comparing different ML models in terms of **effectiveness, efficiency, and stability**, which is the best classifier?

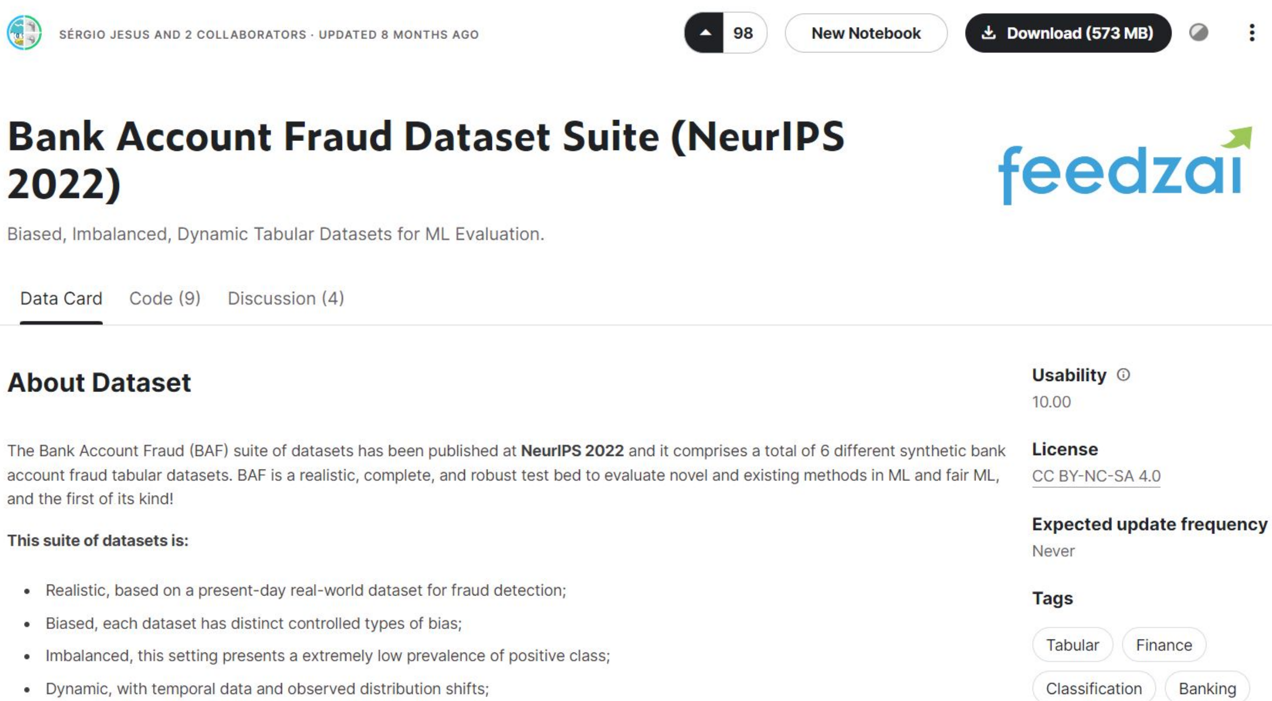
Ryerson University

Test 3 traditional ML models

- It's easy to throw random ML techniques at data.
- But it's harder to understand the business motivations, technical requirements, and stakeholders concerns and find the Mr. Right ML solution.
- This project just test 3 ML (Logistic Regression, Decision Tree, and Random Forest) which learned in the program.

Data Sources

https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022



SÉRGIO JESUS AND 2 COLLABORATORS · UPDATED 8 MONTHS AGO

98 New Notebook Download (573 MB)

Bank Account Fraud Dataset Suite (NeurIPS 2022)

feedzai

Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation.

Data Card Code (9) Discussion (4)

About Dataset

The Bank Account Fraud (BAF) suite of datasets has been published at **NeurIPS 2022** and it comprises a total of 6 different synthetic bank account fraud tabular datasets. BAF is a realistic, complete, and robust test bed to evaluate novel and existing methods in ML and fair ML, and the first of its kind!

This suite of datasets is:

- Realistic, based on a present-day real-world dataset for fraud detection;
- Biased, each dataset has distinct controlled types of bias;
- Imbalanced, this setting presents a extremely low prevalence of positive class;
- Dynamic, with temporal data and observed distribution shifts;

Usability
10.00

License
CC BY-NC-SA 4.0

Expected update frequency
Never

Tags
Tabular Finance Classification Banking

Ryerson University

Employed Methodology

1. Clarity the problem
2. Clarify Constraints
3. Establish Metrics
4. Understand the data sources
5. Explore the data
6. **Clean the data**: Remove, Imputation
7. **Feature selection**
8. **Feature Engineering**: Transformations, Standardize data, One-hot encoding
9. Model selection
10. Model training
11. Model evaluation
12. Deployment

Ryerson University

Clarify Constraints

1. Instances have an order from month 0 to month 7, which means the latter month is dependent to the former month.
2. Assume it’s not cycle or seasonal data, which means it has start month and last month, no cyclical behavior. Month 0 can influence month 1, but month 7 cannot influence month 0. Month 7 is the end.

Measure Matrices

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. ROC-AUC
6. MCC (Matthews Correlation Coefficient)

Clean the data

	Feature Nmae	Missing Rate	Action
1	prev_address_months_count	71%	Drop
2	intended_balcon_amount	74%	Drop
3	current_address_months_count	0.4%	Impute with median
4	bank_months_count	25%	Impute with median
5	session_length_in_minutes	0.2%	Impute with median
6	device_distinct_emails_8w	0.04%	Impute with mode

Convert Wrong Data Type

	Feature Name	Before	After
1	email_is_free	numerical	categorical
2	phone_home_valid	numerical	categorical
3	phone_mobile_valid	numerical	categorical
4	has_other_cards	numerical	categorical
5	foreign_request	numerical	categorical
6	keep_alive_session	numerical	categorical
7	device_distinct_emails_8w	numerical	categorical

Ryerson University

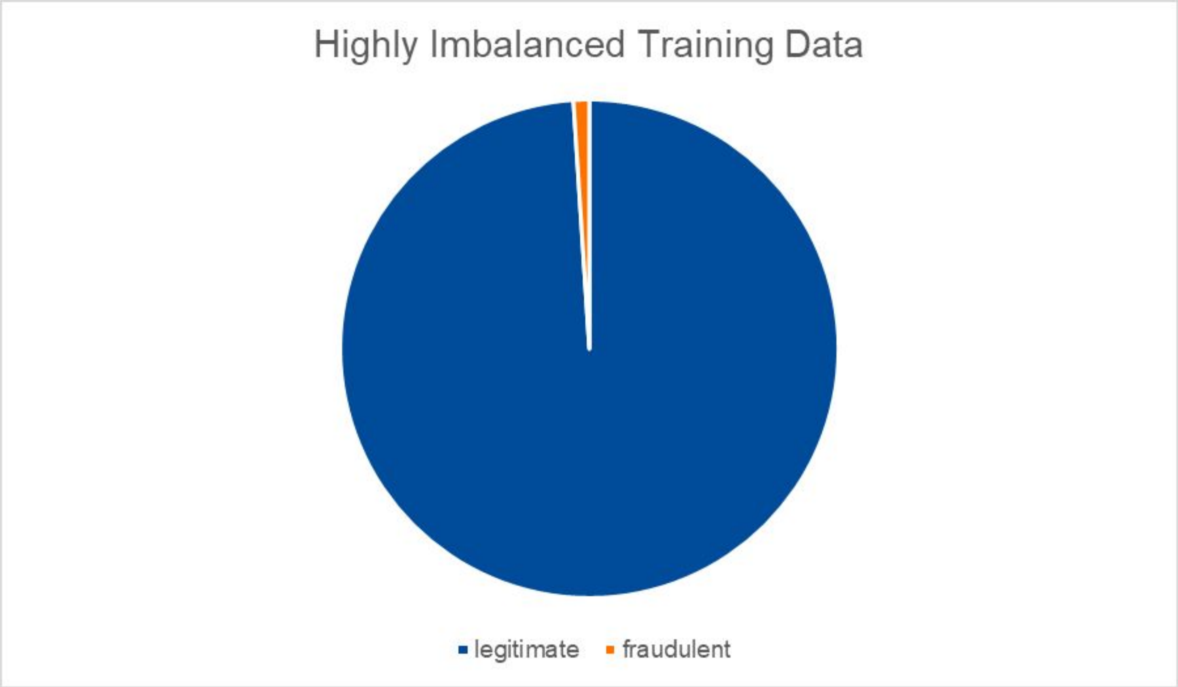
Transformation and Scaling

	Feature name
1	income
2	current_address_months_count
3	days_since_request
4	zip_count_4w
5	velocity_6h
6	bank_branch_count_8w
7	date_of_birth_distinct_emails_4w
8	session_length_in_minutes

One-hot Encoding

After applied the one-hot encoding, the training attributes increased to 57 columns.

SMOTE



SMOTE	Before	After
Class 1: Fraudulent	8151	786,838
Class 0: legitimate	786,838	786,838

Ryerson University

Feature Selection

By analyzing the correlation matrix, ‘elocity_4w’ and ‘month’ are highly correlated with each other. So, I removed ‘velocity_4w’ attribute, this step can improve models’ performance and reduce overfitting.

Ryerson University

Modeling and Evaluation

The workflow I've taken for this part is:

1. Time-series cross validation
2. Dev set evaluation
3. Consistency between the test data and training data
4. Test set evaluation

The three traditional machine learning models I’ve taken are:

1. Logistic regression
2. Decision tree
3. Random forest

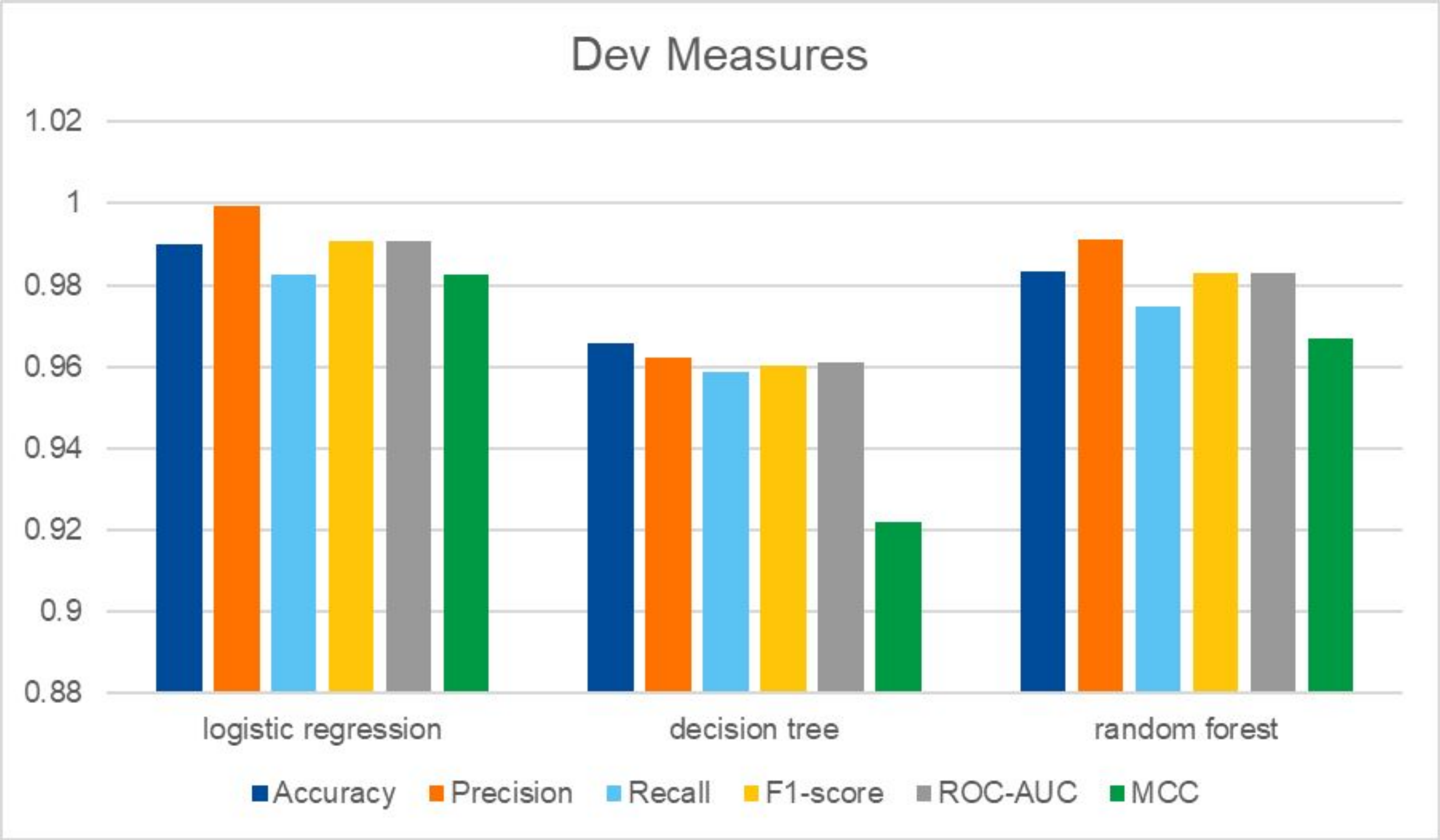
Ryerson University

Iteration

Iteration: i (month)	Time series train data	Time series test data
0	0	1
1	0+1	2
2	0+1+2	3
3	0+1+2+3	4
4	0+1+2+3+4	5

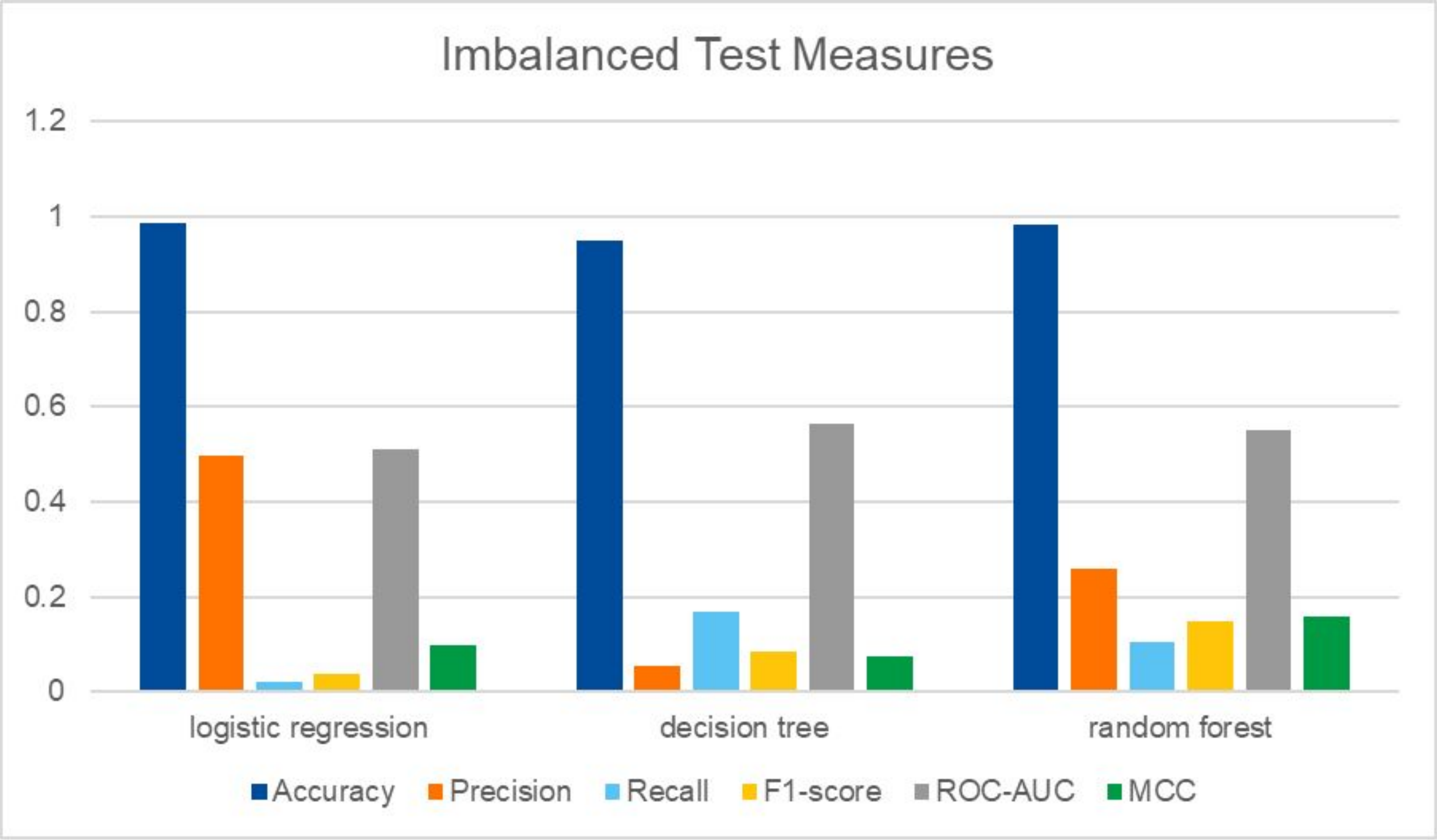
Ryerson University

Effectiveness



Ryerson University

Effectiveness



Ryerson University

Conclusion

1. Traditional machine learning such as logistic regression, decision tree, and random forest are not good at predicting highly imbalanced data, and we need other models to predict such data.
2. If classes are balanced, logistic regression, decision tree, and random forest models can be quite good classifiers.
3. When dealing with temporal data, time-series cross validation is very helpful. It also prevent from overfitting.
4. If features are heavy right skewed, we should transform to shrink the tail first.
5. Feature scaling helps us to standardize the numerical features and make the model more effective.
6. When dealing with time series data, we still need to consider temporal characteristics even in the imputation of missing values step, and never use future data to calculate averages like mean or median.

Ryerson University

Thank you

Ryerson University