



Bank Account Fraud



Name: André Silva
N: 201906559

Fraud Detection


Site: <https://asilva.luxcorp.pt>
Github: <https://github.com/mastersilvapt>
Kaggle: <https://www.kaggle.com/mastersilvapt>

Bank Account Fraud – Problem

- **Highly unbalanced dataset**
- **False Positive better than False Negative**
- **Missing values**
- **Median dense dataset (32 features)**

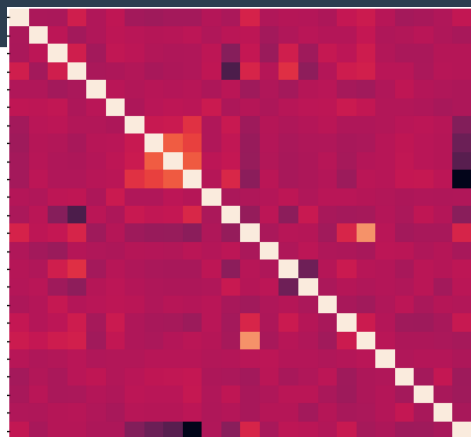


Bank Account Fraud – Data Understanding

- **No middle range incomes**
 - **Fraudulent applications have higher address month count**
 - **More frauds in applications between 30 and 50 years old**
 - **Credit Risk higher on frauds**
 - **BA more common in housing**
 - **Most frauds occur on Windows**
 - **Higher number of frauds in some months**
- 

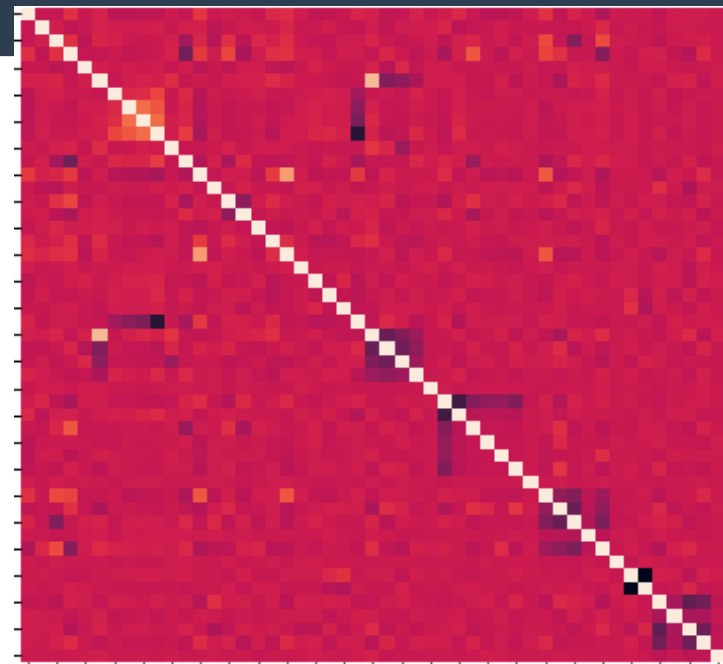


Bank Account Fraud – Data Understanding



RAW

- **Frauds vs Non-Frauds: 1/100**
- **Excessive/Unused/Bad features**
- **“Golden” feature – correlation feature-target**
- **Redundant features – correlation feature-feature**



After Processing

Bank Account Fraud – Data Preparation

- 1st attempt:
- Replace -1 for NAN
- Remove columns with % NAN above 50%, 60%, 70%
- Drop every other record with NAN
- Test data without removing the NAN??

Bank Account Fraud – Data Preparation

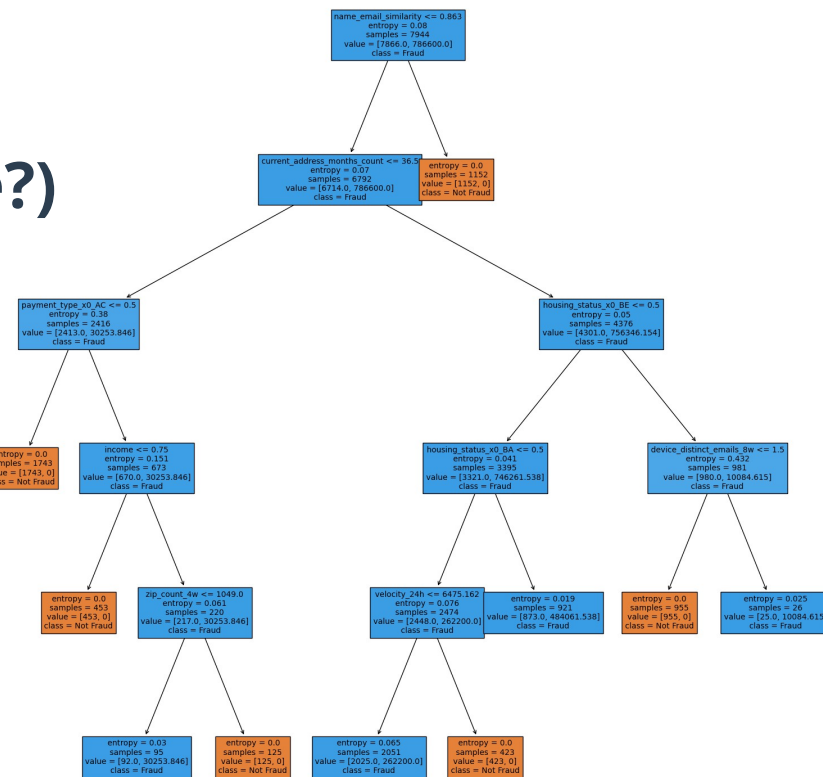
- **2nd attempt:**
- **Iterative Imputer (tricky)**
- **Random Forest Regressor and Classifier**
- **One Hot Encoding**

Bank Account Fraud – Data Preparation

- Split 70% 30% ?
- Little Data!
- Unbalanced Dataset!
- Split into Folds
- Repeated Stratified K Fold
- Split 6 months, 2 months?

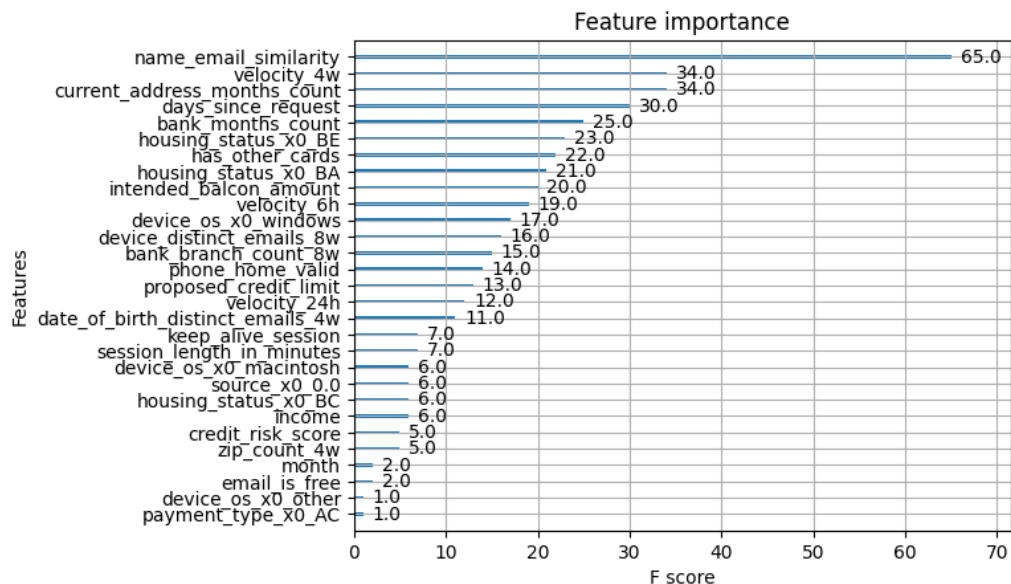
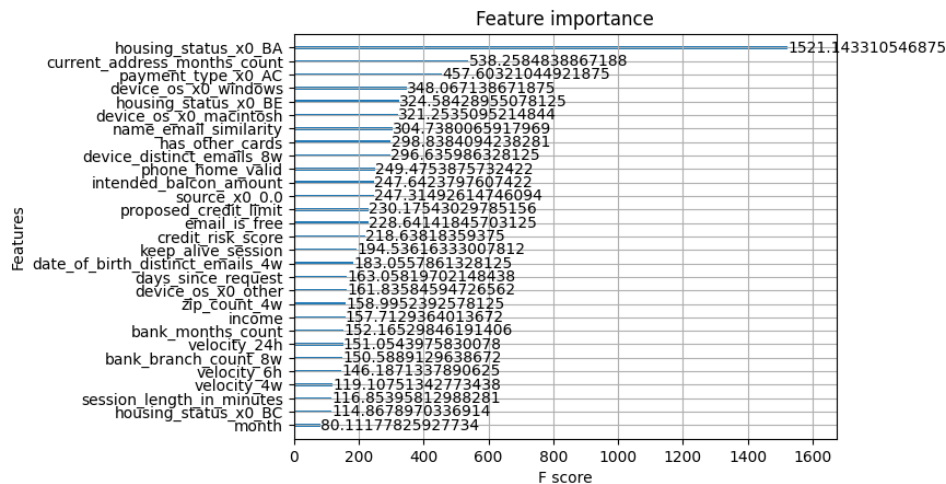
Bank Account Fraud – Model Decision Tree

- Not the best fit!
- Low score (Why submit on kaggle?)
- Grid Search
- Class_weight



Bank Account Fraud – Model XGBoost

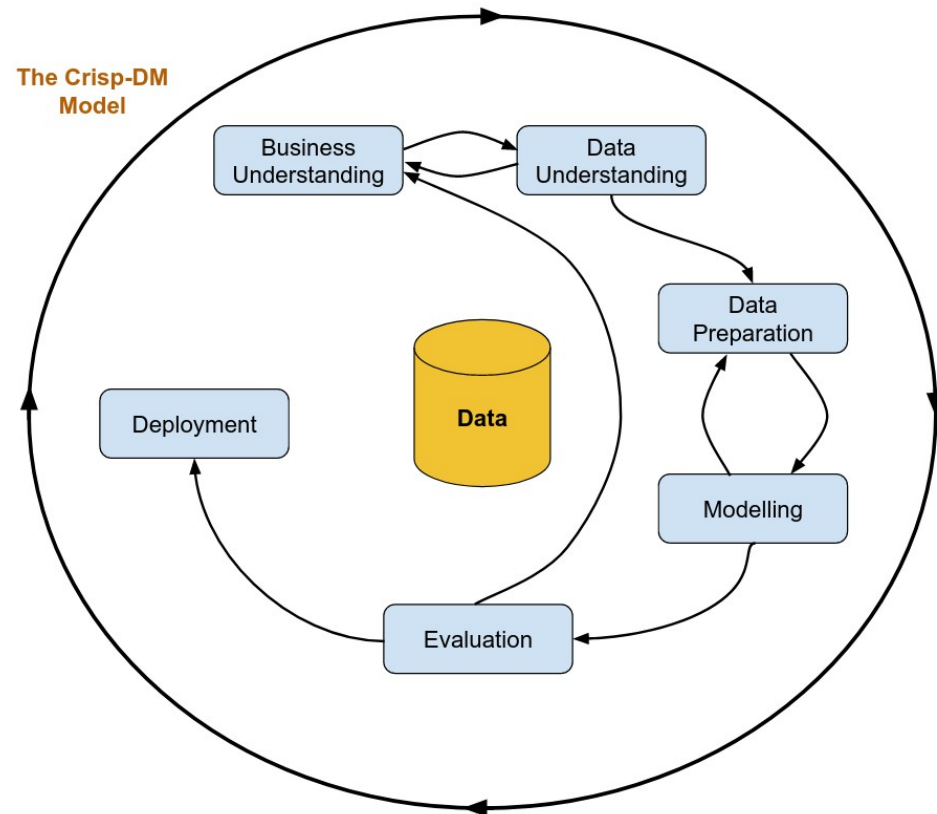
- Grid Search
- Tuning hyper parameters
- Over-fitting?



```
{'eta': 0.05,
'gamma': 0,
'lambda': 5,
'max_depth': 2,
'min_child_weight': 0.6,
'n_estimators': 150,
'scale_pos_weight': 100.84615384615384}
```

Bank Account Fraud – Conclusions

- **Kaggle: 0.84635**
- **More data?**
- **Better Pre-processing!**
- **Another algorithm?**
- **Unsupervised Learning?**
- **Neural Networks?**
- **Kaggle (10 submissions/day)?**





Bank Account Fraud



Name: André Silva
N: 201906559

Fraud Detection

Site: <https://asilva.luxcorp.pt>
Github: <https://github.com/mastersilvapt>
Kaggle: <https://www.kaggle.com/mastersilvapt>