Detecting Fraudulent Job Postings

By Vincent Kwan (301147654)

Advisors:
Prof. Will Au, Prof. David Parent

Table of contents

01. Introduction

02. Analytics Framework

03. Modelling

04. Insights

05. Recommendations

01.

Introduction

Problem Statement, Executive Summary

Fake job offers

- Recent graduate looking for jobs
- Found the perfect offer online
- Pay + Benefits = too good to be true
- Work from Home + Cash payment

2019: 14 million victims | $2 Billion in Losses
2020: Complaints to Canadian Anti Fraud Centre Doubled

Job Fraud is on the Rise

Quick Statistics

$500

Median financial loss due to fake job scams

25 - 35

Age group most vulnerable

62%

Majority of the victims are women

32 hours

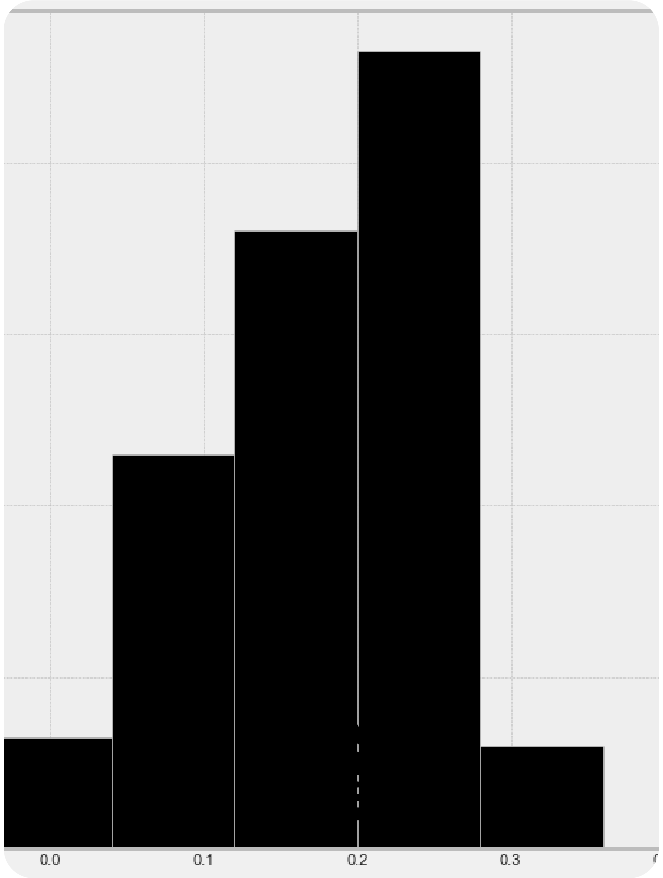
Average time wasted by job seekers on scams

Objectives

1. Critically examine job postings to detect fraud
2. Use a variety of methods & models to detect patterns in the data
3. Gain insight into fraudster behavior
4. Create an analytics framework for problem
    - Data pre-processing and treatment methods
    - Identify novel ways of tackling problem
5. Test common stereotypes about job frauds
    - Poor grammar
    - Clickbait words
    - Cash + Work from Home

Executive Summary

1. Post/Text Characteristics may provide enough predictive power to distinguish between fraud vs non-fraud.
2. Whitespace, consecutive punctuation & Clickbait Ratio highly important features.
3. Scammers target desperate & vulnerable people.
4. Tree-based models had surprisingly high accuracy.



02.

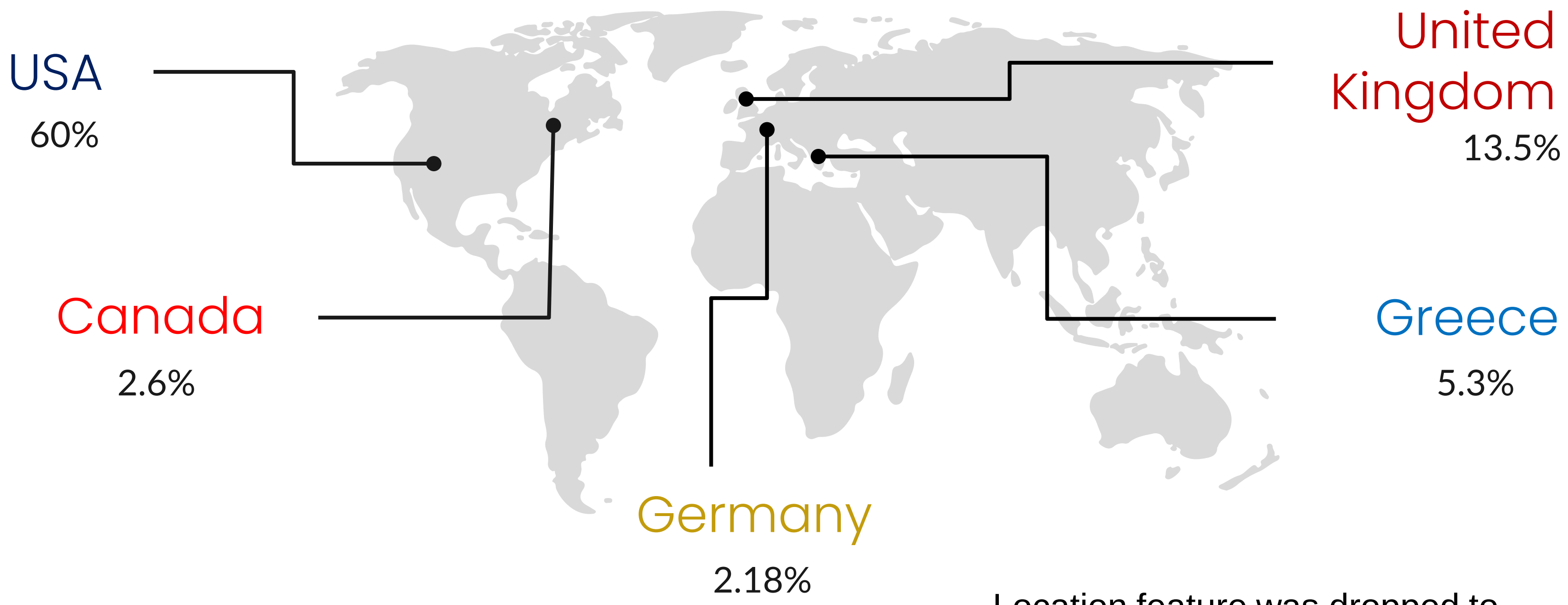
Analytics Framework

Data processing, Methodology, Feature Engineering

Dataset Description

- Employment Scam Aegean Dataset (EMSCAD)
- Obtained from UCI Machine Learning Repository
- Published in 2017
- Contains job posts with various text and categorical fields
- 17,880 records
- 17 fields (including target)
- Binary target - Fraudulent

Dataset Description

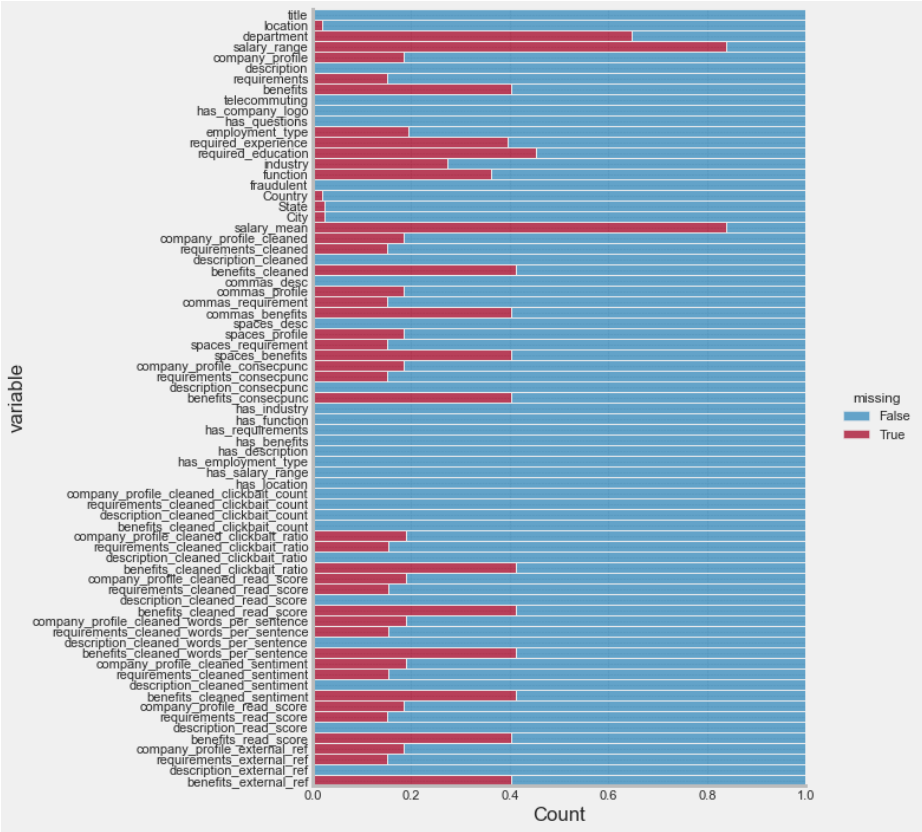


Location feature was dropped to anonymize data.

Missing Data

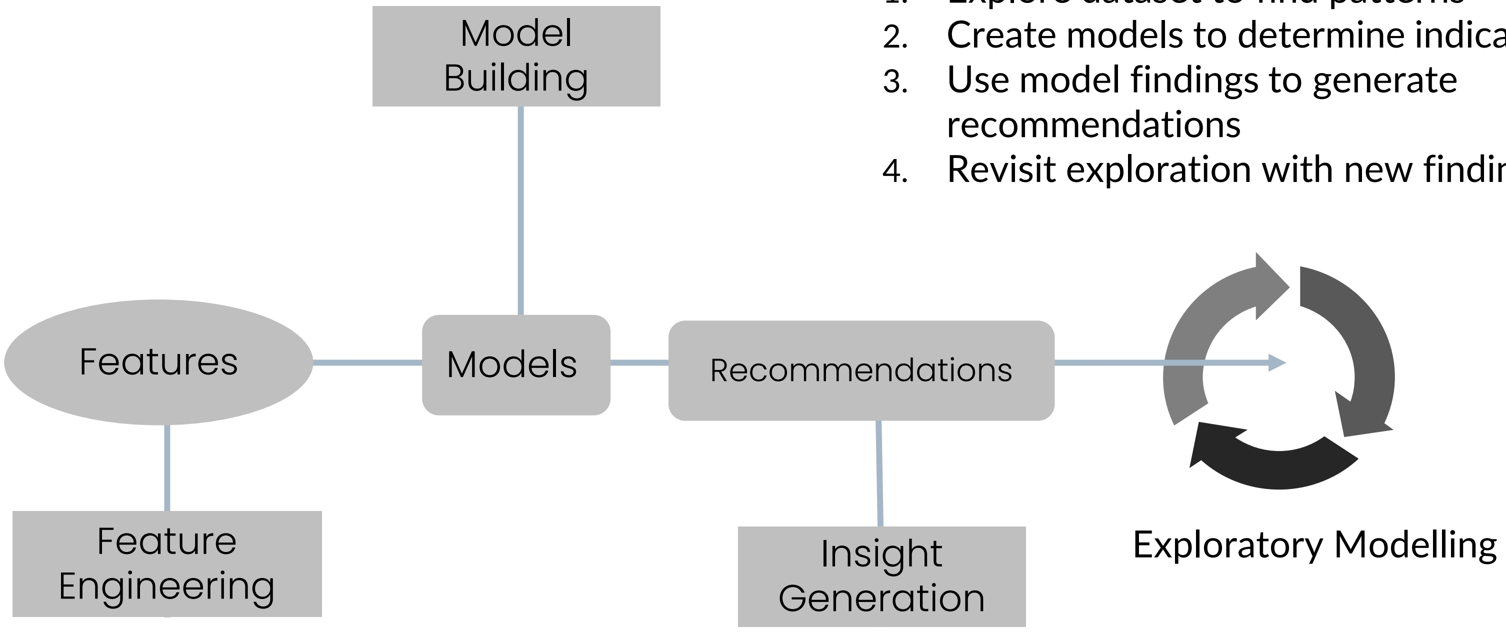
Iterative Imputer used to impute missing data from nearby features.

Features with more than 60% missing data dropped.

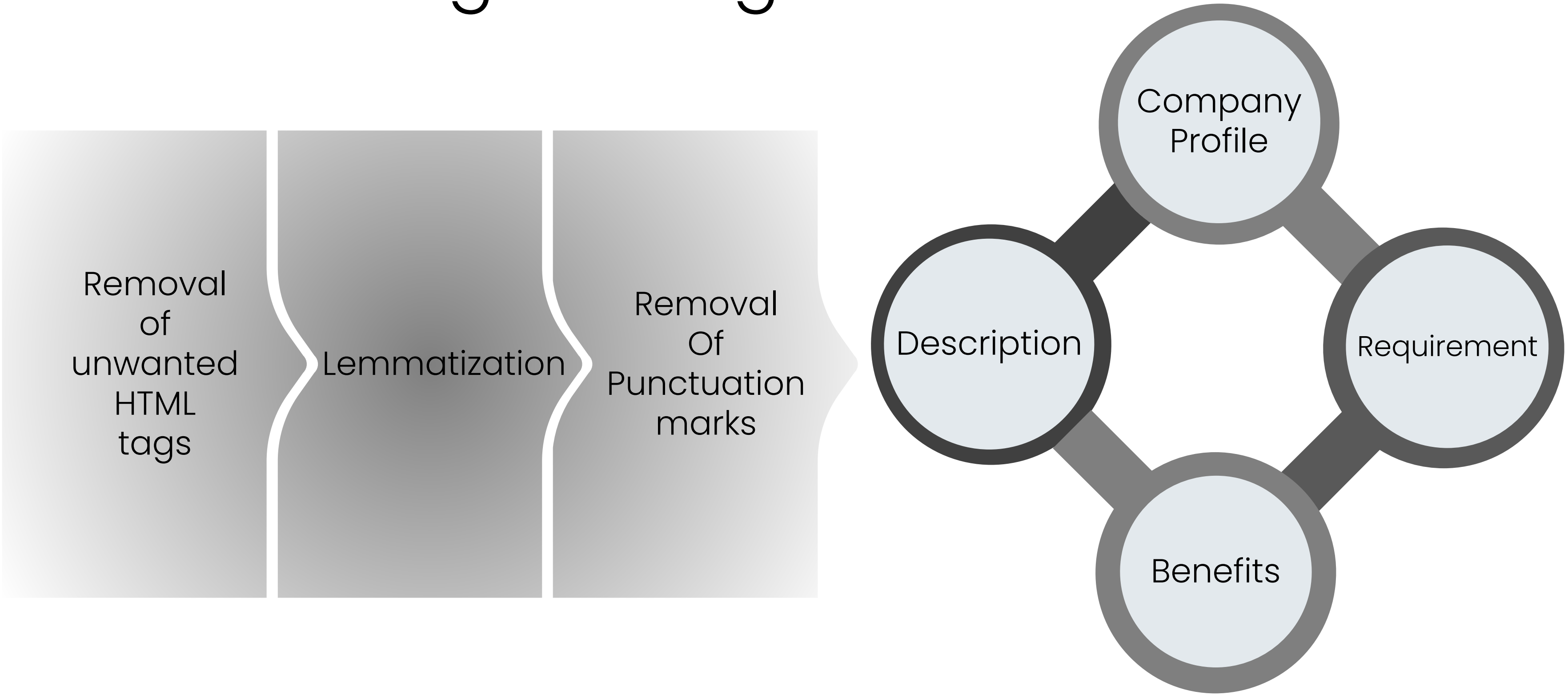


Methodology

1. Explore dataset to find patterns
2. Create models to determine indicators
3. Use model findings to generate recommendations
4. Revisit exploration with new findings



Feature Engineering



Feature Engineering

Post Characteristics

Consecutive Punctuation
Whitespaces
Presence of external references
Clickbait Ratio

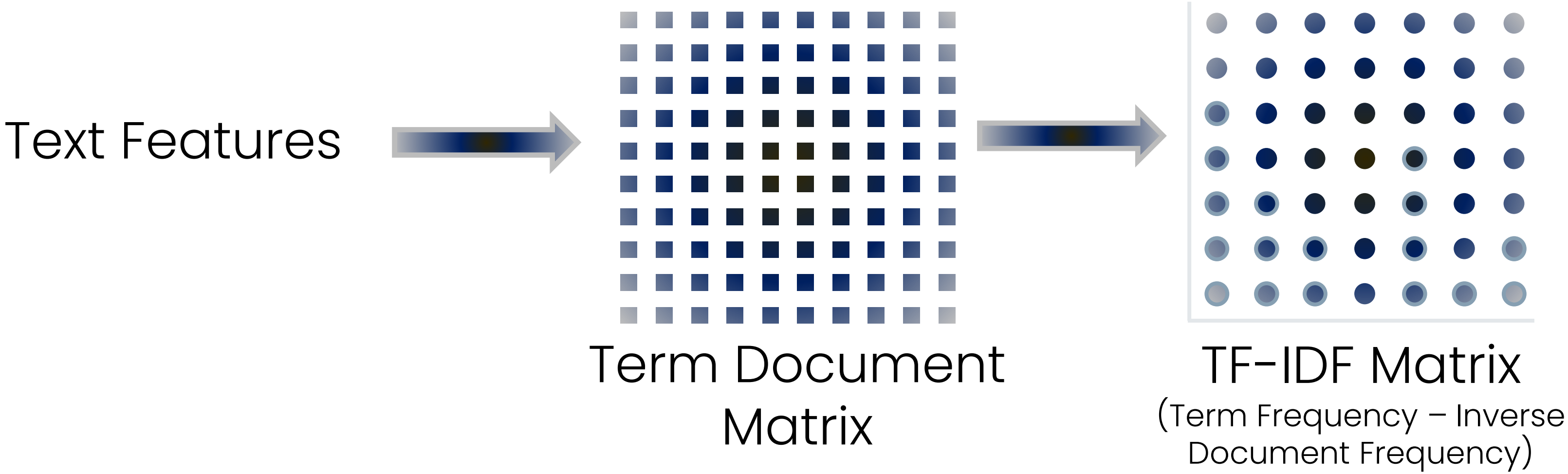
Sentiment

Differences in tone

Reading Score

Flesch reading score used to determine readability

Feature Engineering



Datasets Created

01. Text Characteristics

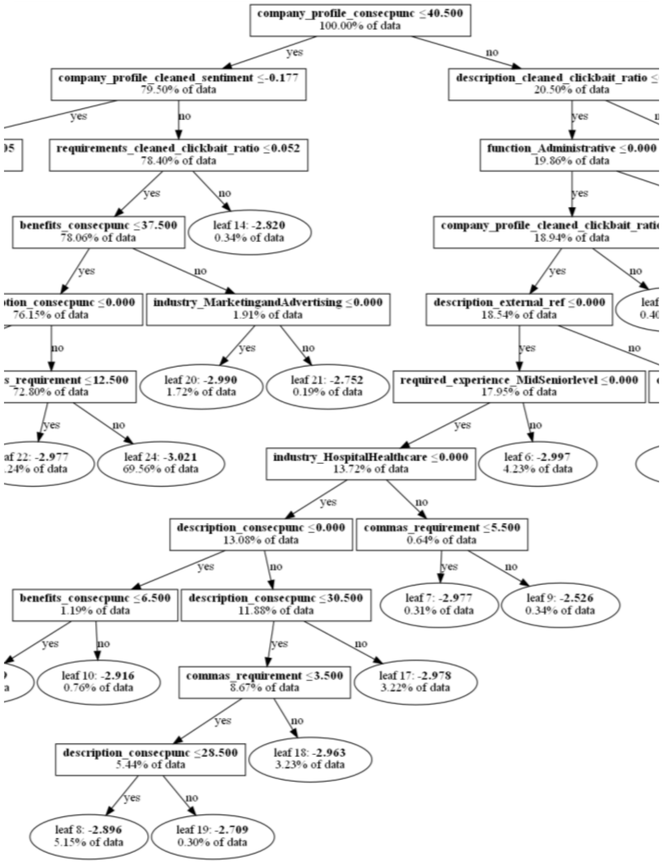
02. TF-IDF Matrix

03. TF-IDF Matrix + Text Characteristics

Two versions of each dataset was created:

1. With upsampling
2. With class imbalance preserved

Datasets were upsampled using SMOTE (Synthetic Minority Oversampling Technique).



03.

Modelling

Modelling Challenges, Inventory of Models, Champion Model

Modelling Challenges

1. **Class Imbalance:** Heavily imbalanced dataset (only 5% minority class) introduces great difficulties in model training and evaluation.
2. **Unwanted elements in data:** Text has unwanted artifacts such as HTML tags, links, formatting metadata, etc.
3. **Large amount of empty fields:** Job posters neglected to fill out all information (e.g. salary range).
4. **Noisy data:** Engineered features were noisy, and didn't convey enough information.
5. **Small dataset:** Only 17.8k records, whereas this problem requires hundreds of thousands if not millions.

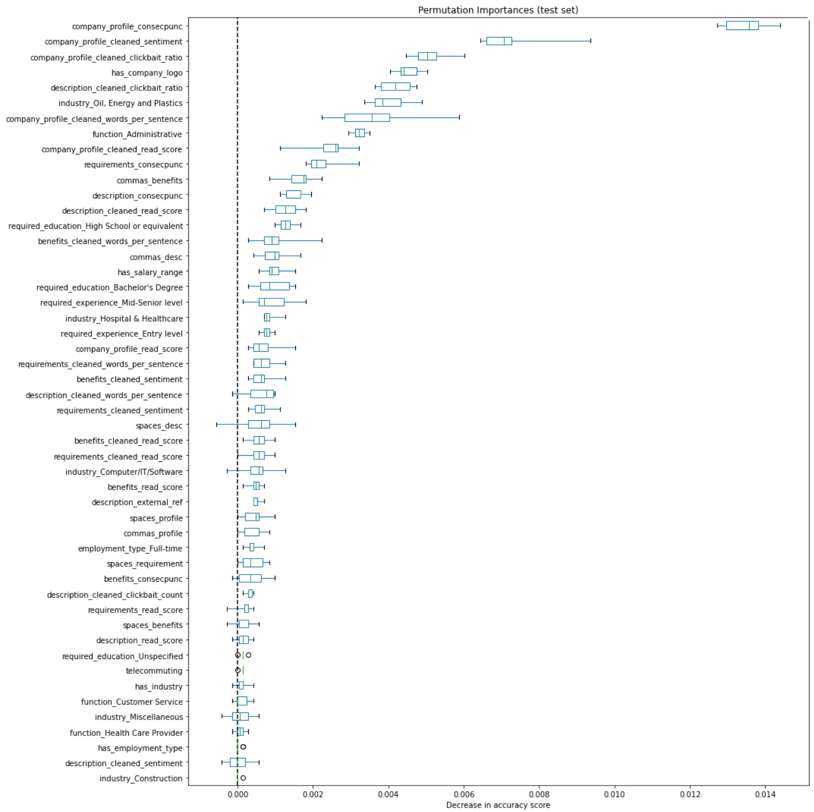
Model Evaluation & Interpretation

F1-Score favoured over Accuracy.

- Naively guessing majority class nets 90% accuracy!
- F1-score only 5%

Permutation Importance used to interpret models.

- This doesn’t necessarily show which feature is most important!
- Shows which features are important for **<u>a given model</u>**



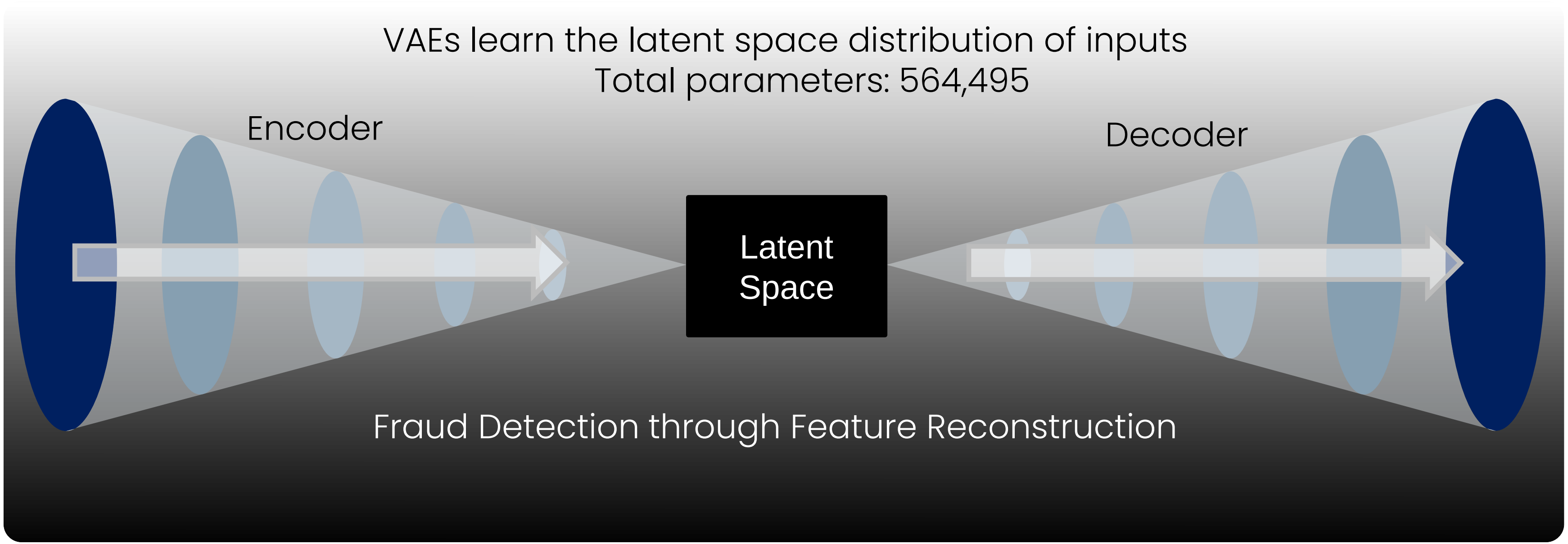
Model Inventory

- Naïve Bayes Classifiers
- Decision Trees + Random Forests
- Gradient Boosted Trees
- Variational Autoencoder

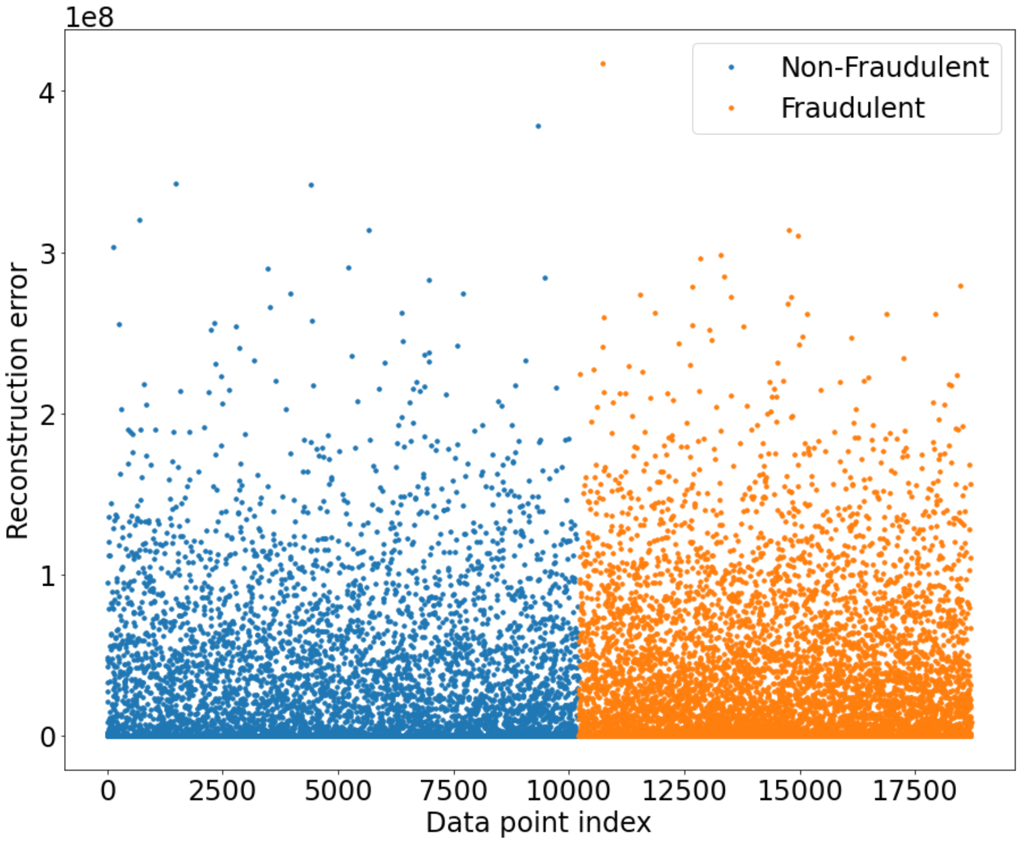
15

Models Created

Variational Autoencoder



Variational Autoencoder



GOAL:

Distinguish between fraudulent & non-fraudulent records through feature reconstruction error.

Generative Adversarial Networks may be better suited to the dataset.

Champion Model

Random Forest

5-Fold Random Search Cross Validation trained only on Text Characteristics

Accuracy

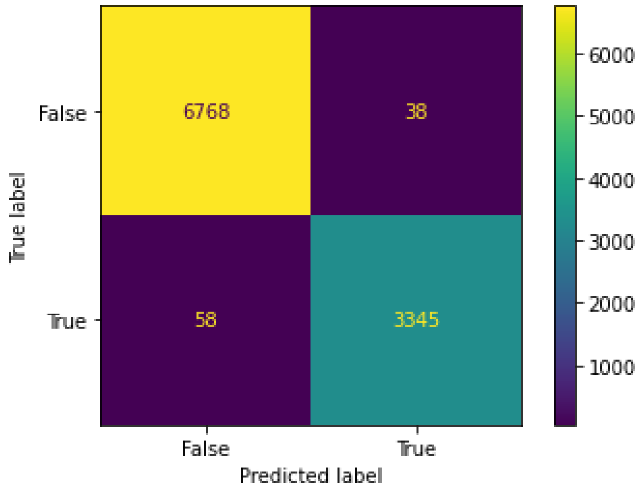
98%

F1-Score

98%

False-Positive Rate

0.55%

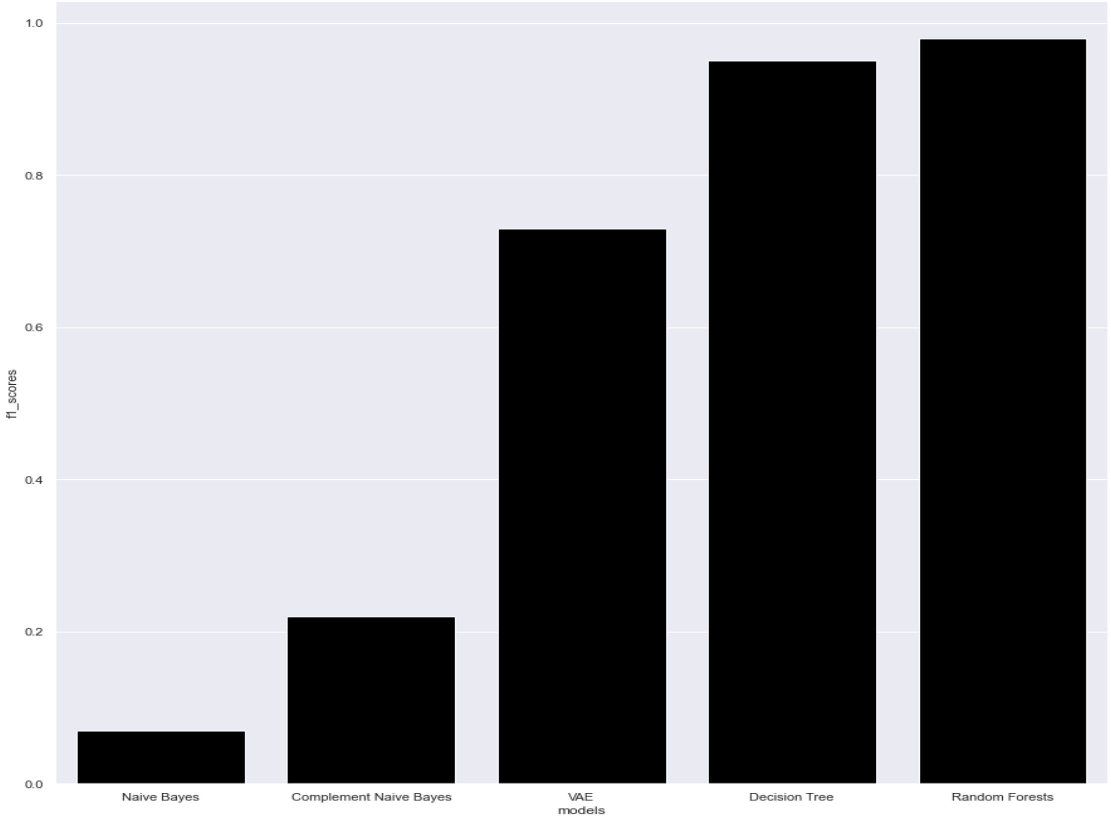


Model Performance

Graph of F1-Scores

Random Forests outperforms all other models.

- Decision Tree has high variance, and fails to perform well for synthetic data.
- Naïve Bayes Classifiers fail in all respects.
- Variational Autoencoders may perform better with more tuning.



04.

Insights

Findings, Heuristics

Main Findings

Whitespaces in text

Most important feature

Oil & IT

These industries have greatest fraud count

Consecutive Punctuation

Higher the score, more likely a fraud

Clickbait Ratio

Higher the score, more likely a fraud

Legitimate businesses prioritized job security in their company profiles

Fraudsters prioritized signing bonuses and perks in their company profiles

Other important indicators of Fraud

High School or Equivalent

Education requirements are lower.

Administrative Roles

Lots of fraudulent offers for Clerk/Secretary roles

Company Logo

Surprisingly, posts with logos have greater proportion of fraud.

Salary Range

Fraudulent posts emphasize monetary rewards over everything else.

05.

Recommendations

Main Recommendations

Continuous Learning

Real-time Analytics system to help detect and prevent fraud.

Protect vulnerable users

Fraudsters clearly have a preference. Protection measures and guidelines should be constructed for the targeted users.

Create a platform for users to share findings

Job Boards must give power back to users – Report & Discussion features should be given greater importance.

At that time, only one type of “robot” truly existed and moved farther and farther in high fields and into dark, ruined rooms. These robots were people.

—Alexander Borovoi, My Chernobyl

Conclusion

Desperate Users are biggest target

Common stereotypes have truth in them

Post attributes may provide enough predictive power

Problem highly suited for Reinforcement Learning

Questions?

References, Code & Statistics available in Documentation