# **Executive Summary**

## **Introduction**

This project was undertaken with the goal of analyzing a dataset consisting of legitimate and fraudulent job advertisement posts. The goal was to seek out Multi-level Marketing posts, instances of corporate identity theft, and ‘stealth’ fraudulent posts (posts masquerading as legitimate offers). However, the end-goal was not to create a model that could discriminate between fraudulent and legitimate posts, but rather to develop a methodology to analyze textual features in job posts, and to derive insights from the data. The Employment Scam Aegean Dataset (EMSCAD) was used for the project – this dataset contains 17,880 real-life job ads collected in the period 2012-2014.

Fake job posts are designed to collect sensitive personal information from job-seekers, or outright scam them of their money. These posts are predatory in nature, and target vulnerable and desperate sections of society. This project has special personal importance to me as I have seen many fraudulent job posts myself.

A novel way of tackling this problem was undertaken – creating a Variational Autoencoder with a compressed encoder in order to learn the latent space distribution of the input features. The learnt distribution was used by the decoder to reconstruct the input features – and the reconstruction loss was used to determine whether the particular job post was fraudulent or legitimate.

I was able to find several indicators of fraud in a job post – most of these indicators were derived from post/text characteristic features (e.g. amount of whitespaces used, consecutive punctuation marks, etc). The main finding was that most fraudulent job posts target lower-educated jobs/professions, including clerk/secretary roles. Fake jobs also prioritize cash rewards/bonuses/perks in their profiles and descriptions as opposed to legitimate job posts prioritizing job security, reputation, challenge and prestige.

## **Executive Objective**

Given a dataset consisting of the details of job postings on a website, detect whether that particular job posting is fraudulent. This involves seeking out multi-level Marketing postings, instances of Corporate Identity fraud, and instances of ‘stealth’ fraudulent posts.

## **Summary of Modelling**

Fifteen models were created in total consisting of two variants of each major class of model. The models consisted of Naïve Bayes Classifiers (Categorical & Complement), Decision Trees & Random Forest Classifiers (fitted with Cross-Validation), Gradient Boosted Trees (Sci-kit Learn’s implementation & LGBM), and finally the Variational Autoencoder.

Extensive data cleansing and preparation was carried out in order to make the data compatible with the models. Feature Engineering was also carried out – creating new features from the text variables. A term-document matrix, and associated TF-IDF scores were created from the text features – this was used with the VAE for classification.

## **Recommendations**

Based on my findings, I recommend:

1. Implementing a Continuous Learning system: The fraud detection problem is especially suited for real-time analytics where statistics and insights can be generated ‘on-the-fly’. Furthermore, a live-model which continuously updates its weights and learning parameters from new data would be highly suited for this problem.
2. Protect Vulnerable Users: Based on my findings, fraudsters target vulnerable and desperate people. They frequently highlight cash rewards, and easy compensation as remuneration. Posts targeting these users, and having similar rewards should be put under greater scrutiny. Potentially suspect posts should be flagged, and users should be informed – similar to social media websites placing a disclaimer on political posts, or posts about COVID-19.
3. Create a platform for Users: Increasingly, discussion forums and Report features have been given less importance by job websites and job boards. These features should be given greater importance in order to give power back to the users. These platforms will allow users to share personal experiences regarding suspicious job advertisements – peer policing will help ensure a higher quality of job posts, and prevent scams.
4. Create a job frauds dataset: The problem of fraudulent job posts is of great importance, considering the monetary and time resources wasted by scammers. Therefore, in order to better study and combat this issue – a job frauds dataset should be created. This dataset would be akin to the ImageNet dataset for image classification. The dataset will allow data scientists, analysts, and programmers to work together in finding significant features to root out fake job ads. Furthermore, this dataset should be continuously updated – allowing users to add suspect posts. This task is particularly labor-intensive, and would require thousands of collaborators to create a labelled dataset.

# **Introduction**

## **Problem Statement**

Fraudulent job posts are job posts masquerading as legitimate job offers with an aim to extract personal & sensitive information from job-seekers, and to scam them of their money.

## **Objectives**

## **Baseline Results**

## **Assumptions & Limitations**

## **Project Dependencies**

## **Dataset Description**

Limitations of the Dataset

# **Data Engineering**

Analytics approach:

1. Bag of words approach: Word based analysis, fraudulent or not based on word frequency

Fraudulent postings are deliberately structured in such a way as to target susceptible and vulnerable people. This can be used to detect these posts.

1. Tone: Sentiment analysis – are fraudulent posts more positive?
2. Post Characteristics and Completeness: Are fraudulent posts incomplete? Do they have more whitespaces or grammatical errors?

## **Text Cleaning & Processing (Lemmatization)**

## **Imputation**

## **Engineered Variables – major classes**

Major classes of variables:

1. Text characteristics: whitespaces, commas used, consecutive punctuation
2. Sentiment and Text Analysis: Fleisch reading score, Sentiment Polarity score
3. Post Completeness: Missing attributes
4. Post Characteristics: employment type, industry, function

## **Upsampling with SMOTE**

# **Modelling**

## **Modelling Approach**

Model Inventory

Kkl;lk;

Evaluation Statistic

## **Naïve Bayes Classifier**

Categorical Naïve Bayes

kjjnkljn

Complement Naives Bayes

## **Tree Models**

Decision Trees

bjkhkjjlkj

Random Forests

## **Gradient Boosted Trees**

Histogram Gradient Boosted Tree

Hyper-parameter tuning

Light Gradient Boosting Machine

Hyper-parameter tuning

## **Neural networks**

Variational Autoencoder

Hyper-parameter tuning

## **Model Comparison**

## **Champion Model**

# **Governance & Control**

## **Data Integrity**

## **Noise in the dataset**

## **Controlling for Drift**

## **A Shifting Target**

## **Variable Level Monitoring**

## **Iterate & Rebuild**

# **Recommendations**

## **Real-time Learning System**

## **Protection Guidelines**

## **Feedback Platform**

## **Job Ad Dataset**

# **Conclusion**

# **Governance & Control**

In any analytics project/undertaking, it is important to outline a Governance & Control procedure acknowledging the limitations, downsides, and risks of using the model and analytical framework put forward by the members of the project. Decision-making with the help of decision-support algorithms and models introduces certain blind-spots and biases into an organization’s decision-making framework. A Data and Model Governance policy helps ameliorate the damages arising from these side-effects, as well as helps to mitigate and prevent systematic errors from creeping into an organization’s decision criteria.

Blind-spots in data and modelling may lead to biased decisions especially when those blind-spots are related to rare events/classes. Insert example of police using model – targeting lower income neighborhoods (Cathy’s book). The risks arising from these blind-spots can be classified into several different categories/tiers. In this project, I will be using the Risk Tiers provided by the European Commission where risks are classified into four distinct classes, namely:

I created a conceptual four-step control procedure for this project. However, given the nature of this project – it will not be possible to establish the fourth and final step of the control procedure. The procedure will consist of:

1. Data Control: In this step, raw data is reviewed, validated, and prepped for feature engineering. Random samples of the raw data are reviewed by a human – the sample size is determined as per the requirements of the project, and size of the dataset.
2. Feature Control: After the raw data has been validated and prepped into a machine-readable/compatible format, existing features are reviewed for skewness, and missing value issues. Feature Engineering is carried out, and variable-level monitoring thresholds are established based on empirical research, and domain knowledge. Variables related to sensitive information like Gender, Age, and Location are monitored or culled as these can lead to bias and fairness issues.
3. Model Control: Once feature engineering has been carried out, modelling assumptions are explicitly listed in order to detect and root out any bias and fairness issues. Appropriate models are chosen based on the analytics problem, and model validation criteria are created. Appropriate metrics to test model performance are chosen.
4. Decision Control: Model results are compared to historical benchmarks, and significant deviations in trends are closely examined in order to ensure robustness of models. Variable-level monitoring is established in order to ensure conformity with modelling environment. Industry experts will be invited to review model results, and critically examine modelling assumptions. It is important to note that these steps are iterative, and not strictly linear. Deficiencies revealed in one step will prompt the entire process to start over again.

## **Variable Level Monitoring**

Highlighted rows are original variables from the dataset. Non-highlighted rows are feature engineered variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Level of Measurement | Python Data Type | Mean | Standard Deviation | Percentage of Positives (True) |
| title |  | object |  |  |  |
| location | Nominal | object |  |  |  |
| department | Nominal | object |  |  |  |
| salary\_range | Ordinal | object |  |  |  |
| company | Nominal | object |  |  |  |
| description |  | object |  |  |  |
| requirements |  | object |  |  |  |
| benefits |  | object |  |  |  |
| telecommuting | Binary | object |  |  |  |
| has\_company\_logo | Binary | object |  |  |  |
| has\_questions | Binary | object |  |  |  |
| employment\_type | Nominal | object |  |  |  |
| required\_experience | Ordinal | object |  |  |  |
| required\_education | Ordinal | object |  |  |  |
| industry | Nominal | object |  |  |  |
| function | Nominal | object |  |  |  |
| fraudulent | Binary | object |  |  | 4.8434% |
| Country | Nominal | object |  |  |  |
| City | Nominal | object |  |  |  |
| State | Nominal | object |  |  |  |
| salary\_mean | Interval | float | 657836.7202 | 22268016.4690 |  |
| company\_profile\_cleaned |  | object |  |  |  |
| requirements\_cleaned |  | object |  |  |  |
| benefits\_cleaned |  | object |  |  |  |
| description\_cleaned |  | object |  |  |  |
| commas\_desc | Ratio | int |  |  |  |
| commas\_profile | Ratio | int |  |  |  |
| commas\_requirement | Ratio | int |  |  |  |
| commas\_benefits | Ratio | int |  |  |  |
| spaces\_desc | Ratio | int |  |  |  |
| spaces\_profile | Ratio | int |  |  |  |
| spaces\_requirement | Ratio | int |  |  |  |
| spaces\_benefits | Ratio | int |  |  |  |
| company\_profile\_consecpunc | Ratio | int |  |  |  |
| requirements\_consecpunc | Ratio | int |  |  |  |
| description\_consecpunc | Ratio | int |  |  |  |
| requirements\_consecpunc | Ratio | int |  |  |  |
| benefits\_consecpunc | Ratio | int |  |  |  |
| has\_industry | Binary | Boolean |  |  |  |
| has\_function | Binary | Boolean |  |  |  |
| has\_requirements | Binary | Boolean |  |  |  |
| has\_benefits | Binary | Boolean |  |  |  |
| has\_description | Binary | Boolean |  |  |  |
| has\_employent\_type | Binary | Boolean |  |  |  |
| has\_salary | Binary | Boolean |  |  |  |
| has\_location | Binary | Boolean |  |  |  |
| company\_profile\_cleaned\_clickbait\_count | Ratio | int |  |  |  |
| requirements\_cleaned\_clickbait\_count | Ratio | int |  |  |  |
| description\_cleaned\_clickbait\_count | Ratio | int |  |  |  |
| benefits\_cleaned\_clickbait\_count | Ratio | int |  |  |  |
| company\_profile\_cleaned\_clickbait\_ratio | Ratio | float |  |  |  |
| requirements\_cleaned\_clickbait\_ratio | Ratio | float |  |  |  |
| description\_cleaned\_clickbait\_ratio | Ratio | float |  |  |  |
| benefits\_cleaned\_clickbait\_ratio | Ratio | float |  |  |  |
| company\_profile\_cleaned\_words\_per\_sentence | Ratio | float |  |  |  |
| requirements\_cleaned\_words\_per\_sentence | Ratio | float |  |  |  |
| description\_cleaned\_ words\_per\_sentence | Ratio | float |  |  |  |
| benefits\_cleaned\_ words\_per\_sentence | Ratio | float |  |  |  |
| company\_profile\_cleaned\_read\_score | Interval | float |  |  |  |
| requirements\_cleaned\_words\_ read\_score | Interval | float |  |  |  |
| description\_cleaned\_read\_score | Interval | float |  |  |  |
| benefits\_cleaned\_ read\_score | Interval | float |  |  |  |
| company\_profile\_cleaned\_sentiment | Interval | float |  |  |  |
| requirements\_cleaned\_words\_sentiment | Interval | float |  |  |  |
| description\_cleaned\_read\_sentiment | Interval | float |  |  |  |
| benefits\_cleaned\_sentiment | Interval | float |  |  |  |
| company\_profile\_external\_ref | Binary | Boolean |  |  |  |
| requirements\_cleaned\_ external\_ref | Binary | Boolean |  |  |  |
| description\_cleaned\_external\_ref | Binary | Boolean |  |  |  |
| benefits\_external\_ref | Binary | Boolean |  |  |  |