# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Fraudulent Job Posting Detection* |
| **Requestor** | *indeed.ca, monster.ca, other job portals (Project is of an academic nature but the results will be practically useful)* |
| **Date of Request** | *06/10/2022* |
| **Target Quarter for Delivery** | *Third Quarter of 2022* |
| **Epic Link(s)** | *N/A* |
| **Business Impact** | *Success of the project will lead to an improved User Experience on job portal websites. The project will also help reveal key characteristics/features of fraudulent job postings which will enable the company to improve its spam detection algorithms.* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

………

**The specific ask:**

*Clearly articulate the specific task you will be conducting to help achieve the opportunity*

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. The project will work with the Employment Scam Aegean Dataset (EMSCAD) which consists of 17,880 real-life job ads collected in the period 2012-2014 which were manually categorized as fraudulent or not.

These fake job postings are designed to extract personal information from unsuspecting applicants in order to target them for advertisements, scams, or to conduct identity fraud. Personal information includes the contact number, email address, physical location, etc. of applicants.

The problem involves analyzing textual data from the job posting contents in order to glean insights into fake job scams. A concern is that scammers will mask their postings to mimic actual job postings – leading to little difference between fraudulent and non-fraudulent posts. While it might seem counter-productive for scammers to include obvious ‘scam markers’ in their postings, it is not a far-fetched idea. For instance, email scams are obvious due to the fact that scammers are looking for non-discerning people who do not notice the red-flags present in the email. This means these people will be more susceptible to their scams (Kane, 2014).

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

The project will build on upon previous research conducted on the dataset, and also utilize general insights given by job portals and job seekers. For instance, creating a ruleset-based classification system can led to high accuracy in detecting fraudulent jobs. The combination of a Gradient Boosting model along with rulesets, and text patterns (bag of words) led to a weighted average F1 score of 88% (Naude, Adebayo and Nanda, 2022). Additionally, in their work, Alghamdy and Algharby (2019) showed that a simple non-ruleset-based classification utilizing Random Forests would lead to an accuracy as high as 97%. However, a potential issue with their work is low Precision and Recall scores for the Truth (Fraudulent) class records.

Apart from research, the project will also test general suggestions and advice for job seekers by experienced professionals and job portals. An analytical approach will be taken to test the veracity of ‘general knowledge’. In a support post, Indeed suggests that most fake/scam jobs are posted for Receptionist and Driver roles. This and many other suggestions/advice will be tested in the project.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

Fake job postings on job portals leads to a poor customer experience for job-seekers and students. But the damage extends beyond a poor experience – people scammed by fake jobs lose time and money. While most discerning people don’t fall for the scams, some people still fall prey. According to the Better Business Bureau (2022), the median financial loss due to fake job scams was $500. Young professionals in the age group of 25 to 35 years old were victimized heavily, and women comprised 62% of the victims. Direct losses due to job scams amounted to $2 billion. In a report released by the Federal Trade Commission (FTC) – in USA, Americans lost $68 million to job scams in 2022 alone. Therefore, job portals like Indeed.ca have a responsibility towards their users/customers to detect and remove fraudulent job posts.

This project will benefit job-seekers and society at large. Removing fake jobs will lead to a much better user experience, as well as help prevent victims from being scammed. It is hard to estimate time and cost savings, however, we can roughly say that since the median financial loss due to job scams is $500, and the minimum wage in Ontario is $15.5 – the project will prevent approximately 32 hours of a job-seeker’s time from being wasted by scams.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

The main objective of the project is to gain insights into fraudster behavior, and cultivate a set of heuristic rules which will allow potential job seekers to avoid job scams.

A secondary objective of the project is to generate a systematic, and methodical process of analyzing job posting in order to glean useful information as to its authenticity, validity, and truthfulness. Related to this aspect is a minor secondary objective in which general truths, advice and suggestions given by professionals and experienced job seekers will be tested in an analytical manner. These ‘general truths’ will be examined critically to determine their veracity.

Finally, the project will also seek to create models that can discern between fraudulent and non-fraudulent job postings in a limited setting. Additionally, an attempt will be made to generate probabilities for whether a job posting is a scam.

The project will also seek to test the hypothesis that: *Fake job postings are obvious (many grammatical and spelling errors, inconsistency in details) so as to attract oblivious candidates who are better targets for scamming.*

Key Questions:

1. Are fraudulent job posts incomplete?
2. Do fraudulent job posts have bad grammar and punctuation?
3. Is there a common “style” to fraudulent job posts?
4. Is there a text feature powerful enough to give us a strong indication of fraud alone?

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

Assumptions:

1. Typos, spelling errors, and mistakes reflect ground truth and are not a result of improper data extraction.
2. Data reflects objective truth on job portals, and has not been doctored.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | Typical classification algorithm success measures will be used such as: Accuracy, F1-score, Precision, Recall, ROC, Calibration plot. More measures will be used if required as per further research. |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** Classification Task – detect whether a record is fraudulent or non-fraudulent.

However, we must note that detecting and removing fake job postings/fraudulent job posts extends beyond normal data analysis. In certain situations, ‘smart scammers’ mask their posts in such a manner that it impossible to determine whether it is a scam from the job posting details alone. As such, this project’s main goal is not simply to create a model to classify job posts as fraudulent, but also to gain insight into the behavior of job scammers. This insight will enable job portals like Indeed.ca to improve their algorithms, and advise their users of how to safeguard themselves from these scams.

In other words, this project extends beyond a simple binary classification task – this project also seeks to gain a deeper knowledge into the characteristics of fake job posts, and behavior of scammers. Thus, this project also involves application of business knowledge & acumen.

**Methodology:** The main approach will consist of Processing, Tokenization, Feature extraction, and Modeling.

First, EDA will be conducted to determine the most common characteristics among fraudulent job postings. Meticulous text processing and cleaning will be conducted to rid the data of unwanted artifacts and elements. For instance, the text field contains elements like <u> to signify HTML elements. These elements will be removed in the text processing step. Then, a data quality check will be conducted to detect error records and duplicates. The data quality check will be done systematically – both automatically, and manually.

After processing, cleaning, EDA, and data quality checking – feature engineering will be conducted. This will include generating new features based on textual patterns such as number of consecutive punctuation marks, blank spaces, or grammatical/spelling errors present. The feature engineering task will be based on research conducted on the dataset (as seen in this [paper](https://link.springer.com/article/10.1007/s00146-022-01469-0#Abs1) published in the AI & Society journal). Feature extraction will then be conducted. Various methods like LDA, and PCA will be applied to determine the most useful features for modeling.

After feature engineering, modeling will be conducted. Models will include but not be limited to Naïve Bayes, Random Forests, Logistic Regression, Artificial Neural Networks, Support Vector Machines, LightGBM and XGBoost.

Finally, the models will be used to build a set of heuristic rules, insights, and recommendations. These final artifacts will be useful not only to job portals like Indeed.ca but also potential job seekers seeking to avoid fraudsters.

**Output:** The main output will be a set of insights and rules which will be built based on the results of modeling and EDA. These insights will then be used to build a set of recommendations for job portals like indeed.ca to apply.

The project will also lead to the creation of a methodical system of analyzing job postings. Job portals will be able to build and improve on the frame of analysis, and methods used in this project in order to further improve their user experience.

Finally, models will be built with the ability to discern between fraudulent and non-fraudulent job posts – but this will be used as a supporting task to gain further insight on fraudster behavior.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** Audience consists of all job seekers searching for jobs on job portals like indeed.ca, monster.ca or LinkedIn

Population consists of all job posts on every job portal

**Observation window:** 2012 to 2014

**Inclusions:** N/A

**Exclusions:** N/A

**Data Sources:** EMSCAD

**Audience Level:** Professional

**Variable Selection:** Prominent text features will be chosen after tokenization, and feature engineering.

**Derived Variables:** Derived variables will be obtained after feature engineering. These will include text markers such as number of punctuation marks used, number of whitespaces, number of mis-spellings, etc.

**Assumptions and data limitations:** Assumptions have been listed in Section 2.1. Data Limitations have been listed in Section 4.0.

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependencies & Risks** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| *Small sample size* | *N/A* | *N/A* | *Small sample size greatly limits the generalizability of the modeling output, and insights derived from dataset* |
| *Changing status quo* | *High* | *N/A* | *Fraudsters often change their behavior in order to bypass spam-detection, and fraud-prevention measures. The changing status quo will mean limited applicability of the resulting models and insights.* |
| *Outdated Dataset* | *N/A* | *N/A* | *The dataset consists of job posts collected in the 2012-2014 period. Therefore, a valid criticism is that the dataset is outdated. However, it must be noted that the main aim of the project is to gain an insight into fraudster posting behavior. This particular characteristic remains unchanged for the most part as evidenced by the unchanging nature of email scams.* |
| *Imbalanced Dataset* | *N/A* | *N/A* | *The data is highly imbalance which leads to problems when modeling. Oversampling has to be conducted in order to counteract this measure.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

**Please note that the dates are tentative, and may be subject to change as project requirements evolve.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request | The project will be formally initiated by sending a request to the advisory team. |  | *5* | *July 15th* |
| 2. | Assessment / Triage | Exploratory analysis will be conducted to determine key/major issues in data quality and feature engineering. Data exploration will also yield insights into the distribution of the data, and target variable characteristics.  Once key characteristics and issues have been identified, these will be triaged and prioritized based on importance. | Limited to exploration, quality check, and high-level discussion. | 7 | *July 22nd* |
| 3. | Prioritization | 7 | *July 22nd* |
| 4. | Data Exploration   * Issues with duplicates * Issues with Spend data | 7 | *July 22nd* |
| 5. | Story Board | Once data exploration has been conducted, a user-story will be created to determine how potential job-seekers are scammed by fake job posts. This will help extract key areas to focus on. | *Limited to user-testing and survey of general user experience of job portals, particularly related to job scams.* | *1* | *July 23th* |
| 6. | QA Output | The data quality issues identified in step 4 will be addressed, and the results of that will be presented to stakeholders. | *Limited to data quality issues prioritized in Step 4.* | *7* | *July 25th* |
| 7. | Data Processing & Feature Engineering | The textual data will be tokenized and processed to remove unwanted elements. Imputed variables will be created, and feature engineering will take place simultaneously. Feature engineering will build on prior work from researchon the dataset. | *No modeling shall take place. Feature engineering will remain blind to modeling in subsequent step so as to ensure validity of analysis, and freedom from biases.* | *4* | *July 29th* |
| 8. | Modeling | Once the data has been processed, it will be used for modeling through classification algorithms.  In this step, the architecture for the Neural Network will be determined through a trial-and-error approach utilizing cross-validation techniques. | *Models used: LightGBM, XGBoost, Logistic Regression, Random Forests, Simple Decision Tree, Artificial Neural Network (architecture to be determined), SVM* | *6* | *August 4th* |
| 9. | Insight Generation & Presentation creation | Heuristic rules and recommendation will be built. The presentation shall be created simultaneously, as results will inform presentation format. |  | *2* | *August 6th* |
| 10. | Advisory Team Presentation | Results will be presented to Advisory team. |  | *1* | *August 7th* |

# References

*Fake job scams on the rise during pandemic, warns the Better Business Bureau*. (2021, September 9). Retrieved from cbc: https://www.cbc.ca/news/canada/british-columbia/fake-job-scams-on-rise-during-pandemic-1.6169860

Kane, L. (2014, May 8). *There's a reason Nigerian scammers are so obvious in their emails*. Retrieved from Business Insider: https://www.businessinsider.com/why-nigerian-scam-emails-are-obvious-2014-5

Liu, J. (2022, June 10). *Americans lost $68 million to job scams this year—here’s what to look out for*. Retrieved from cnbc: https://www.cnbc.com/2022/06/10/americans-lost-68-million-to-job-scams-this-year-here-are-the-most-common-ones.html

Naudé, M., Adebayo, K.J. & Nanda, R. *(2022)A machine learning approach to detecting fraudulent job types.* AI & Soc.doi: 10.1007/s00146-022-01469-0

Alghamdi, B. and Alharby, F. (2019) *An Intelligent Model for Online Recruitment Fraud Detection.* Journal of Information Security. doi: [10.4236/jis.2019.103009](https://doi.org/10.4236/jis.2019.103009).