

# UNIVERSITY OF CONNECTICUT

**OPIM 5671- DATA MINING AND BUSINESS INTELLIGENCE**

Sudip Bhattacharjee

# Text Mining Group Project

**Decoding Deception: Unveiling Fraudulent Job Postings**

Madhavi Muppani

**Table of Contents**

1. Executive Summary
2. Business Case
3. Dataset Preparation
   1. Removing Null Values
   2. Rejecting Unnecessary Variables
   3. Concatenating Text Variables
   4. Preprocessing the Data
4. Model Development
5. Model Evaluation
   1. Performance Metrics
   2. Title + Description as Text Field
   3. Company Profile as Text Field
   4. Description as Text Field
6. Conclusion & Recommendations
   1. Recommendations for Job Seekers
   2. Recommendations for Job Sites
7. References

**1) EXECUTIVE SUMMARY**

As MSBAPM students prepare to enter the job market, the risk of encountering fraudulent job postings is a growing concern. These postings often aim to steal personal information from job seekers, posing a significant threat to privacy and security of students. To address this issue, performed a Text Mining analysis of fraudulent job postings to identify common patterns and keywords used in these postings. By understanding these patterns, goal is to develop recommendations for peers to detect and prevent unnecessary application time and potentially insecure outcomes.

Explored a dataset from Kaggle that contains 18,000 rows of job postings labeled as either legitimate or fake, with 18 various attributes. After preprocessing the data, I developed various models using Decision Trees and Logistic Regressions, focusing on different text fields for each permutation. I evaluated these models with accuracy measures and found that the Log and Entropy Decision Tree with high resolution and a high SVD dimension of 120 produced the best accuracy of 82%.

Based on analysis, I recommend job seekers to pay special attention to the company profiles when evaluating job postings. They should watch out for caution words such as "signing bonus," "perks," and "career growth," which are commonly used in fraudulent postings. Job sites should also enroll in Quality Control Teams to assess job postings thoroughly before posting them on their Ibsite. These teams can proactively flag potential fraudulent postings and ensure a safer job search experience for users.

**2) BUSINESS CASE**

I feel obligated to protect peers from fraudulent job postings that aim to steal personal information from Ill-intentioned young adults. According to the [Society for Human Resce Management](https://www.shrm.org/topics-tools/news/beware-fraudulent-job-postings-aim-to-steal-money-identity#:~:text=Fraudsters%20posing%20as%20recruiters%20and%20employers%20are%20contacting,extract%20Social%20Security%20numbers%20and%20other%20personal%20data.), there is an increasing risk of individuals and companies posing as employers to extract Social Security numbers and other sensitive data, posing a significant threat to privacy and security. In fact, over 183,000 frauds Ire reported to the [FTC](https://www.ftc.gov/business-guidance/blog/2023/08/job-scammers-go-even-lower-way-they-hire?tpcc=NL_Marketing) in 2023 alone. Every individual who was affected by fraudulent job postings lost $900 on average, due to the financial implications of sharing personal information with scammers ([SHRM](https://www.shrm.org/topics-tools/news/beware-fraudulent-job-postings-aim-to-steal-money-identity#:~:text=Fraudsters%20posing%20as%20recruiters%20and%20employers%20are%20contacting,extract%20Social%20Security%20numbers%20and%20other%20personal%20data.)).

To address this issue, I are proposing a Text Mining analysis of fraudulent job postings to identify common patterns and keywords used in these postings. By understanding these patterns, goal is to develop recommendations for peers to detect and prevent unnecessary application time and potentially insecure outcomes. I strongly feel that analysis can help to make a difference and protect fellow classmates as they kickstart their careers as change-making young professionals in the business analytics profession.

**3) DATASET PREPARATION**

Explored a dataset from Kaggle titled "[Real or Fake: Fake Job Posting Prediction](https://usc-word-edit.officeapps.live.com/we/The%20dataset%20provided%20on%20Kaggle%20titled%20%22Real%20or%20Fake:%20Fake%20JobPosting%20Prediction%22%20is%20ideal%20for%20this%20project%20as%20it%20contains%20a%20comprehensive%20collection%20of%20job%20postings%20labeled%20as%20either%20real%20or%20fake.%20This%20dataset%20includes%20various%20features%20such%20as%20job%20description,%20company%20profile,%20and%20required%20qualifications,%20which%20are%20crucial%20for%20text%20mining%20analysis.%20By%20utilizing%20this%20dataset,%20we%20can%20train%20our%20machine%20learning%20model%20to%20accurately%20distinguish%20between%20legitimate%20and%20fraudulent%20job%20postings%20based%20on%20the%20identified%20patterns%20and%20keywords.%20Additionally,%20the%20availability%20of%20a%20labeled%20dataset%20streamlines%20the%20model%20development%20process,%20allowing%20for%20efficient%20training%20and%20evaluation%20of%20the%20model's%20effectiveness%20in%20detecting%20fraudulent%20job%20postings.%20Overall,%20the%20use%20of%20this%20dataset%20enhances%20the%20robustness%20and%20reliability%20of%20our%20analysis,%20contributing%20to%20the%20project's%20success%20in%20protecting%20job%20seekers%20from%20fraudulent%20activities.).” The data contains a comprehensive collection of job postings labeled as either legitimate or fake. This dataset includes 18,000 rows of job postings and 18 various attributes such as:

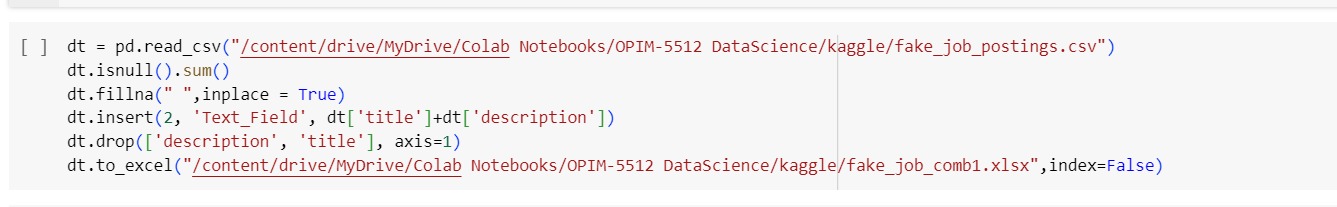
|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| job\_id | Unique identifier for each job posting | Integer |
| title | Title of the job | String |
| location | Location of the job | String |
| department | Department within the company to which the job belongs | String |
| salary\_range | Salary range for the job | String |
| company\_profile | Brief description of the company offering the job | String |
| description | Detailed description of the job | String |
| requirements | Requirements or qualifications for the job | String |
| benefits | Benefits offered with the job | String |
| telecommuting | Indicates whether the job allows telecommuting | Binary |
| has\_company\_logo | Indicates whether the company posting the job has a logo | Binary |
| has\_questions | Indicates whether the job application includes questions for the applicant | Binary |
| employment\_type | Type of employment | String |
| required\_experience | Required level of experience for the job | String |
| required\_education | Required level of education for the job | String |
| industry | Industry to which the job belongs | String |
| function | Function or role of the job within the company | String |
| fraudulent\* | Indicates whether the job post is fraudulent | Binary |

*\*Target Variable*

Initial exploration of the data confirmed that only 4% of the 18,000 rows contained valid data from every variable. I also discovered severe skewness in data, with 95.2% of job postings categorized as real. This sparked attention, with noting the importance of variable selection and stratification to ensure that models are not biased towards the majority class and can effectively capture patterns in the fraudulent data.

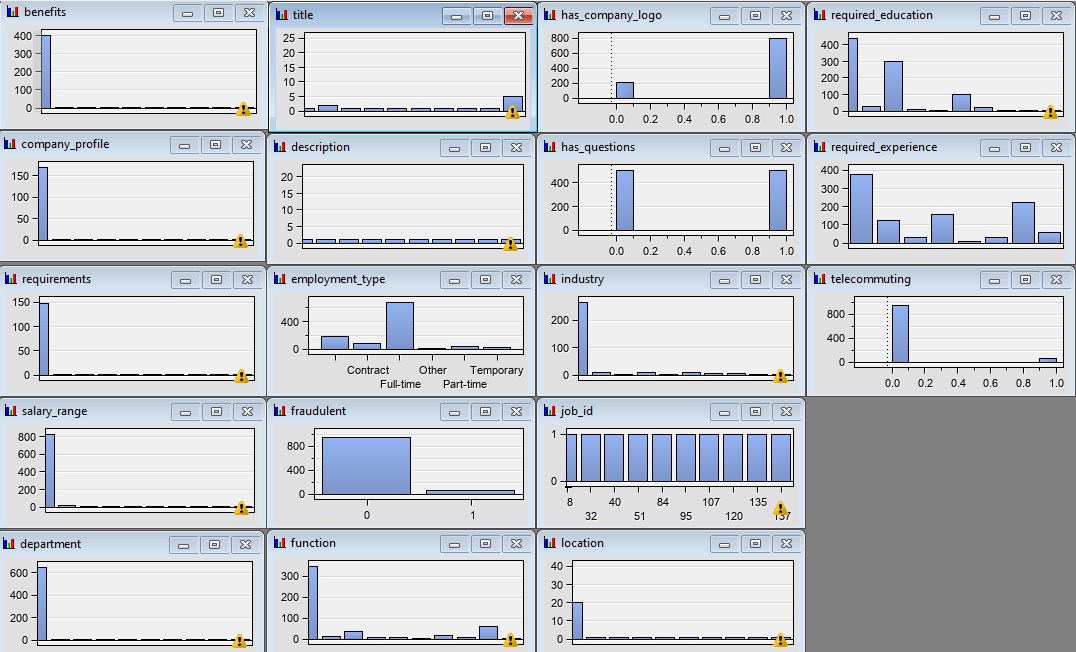
**3.A) Removing Null Values**

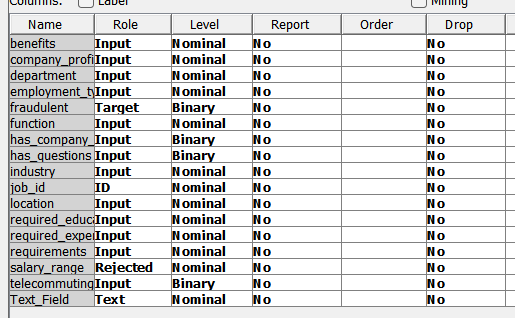
I used the code below for preprocessing the data, looked for any null values in the dataset that could be removed. Upon removing those values, I determined there are no duplicate values that could lead to inaccurate calculations. Otherwise, such repeated data may create confusion and make it difficult to identify and analyze unique data points.



**3.B) Rejecting Unnecessary Variables**

To help streamline inputs into the modeling work, I identified several variables that would add no value to the analysis. The screenshots below help us to illustrate how the Department, Salary and Industry provide highly skeId data. At initial glance, it also seemed like Company Profile, Title and Description may also provide minimal value, since the distributions are either highly skeId or evenly dispersed across documents. HoIver, wanted to explore the text fields further to determine whether it was in best interest to remain in the confinements of the data as made available to us, or whether there would be a valuable opportunity to combine text fields as a new, unique variable for consideration.





*Important Note: The Text\_Field variable is a field concatenating several existing text variables. This is discussed in greater detail later in this paper.*

**3.C) Concatenating Text Variables**

In the real world, if someone started searching online, it could be tremendously difficult for them to analyze multiple text fields with equity across job postings that may or may not provide information for each. I quickly learned the importance of building a model with only one text variable, particularly because of how rare it is to find a job posting with all 18 variables. By combining information from multiple text fields, I simplified analysis without sacrificing the context that fields help to provide. goal was to therefore help ensure model is leveraging as much information as possible in analysis, without adding any more complexity that would be otherwise difficult to explain to the “customer.”

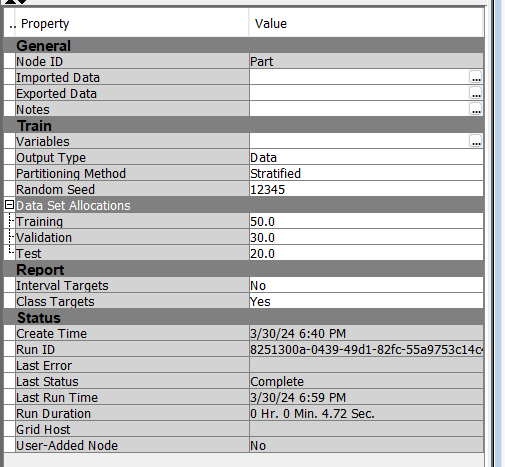
There Ire three text variables that stood out to team as providing value to the model’s performance: Title, Description and Company Profile. Both Description and Company Profile include more text than just the job title, so I Ire inclined to keep both variables as standalone options to build models on. HoIver, concatenating Title and Description would also help to provide additional context to the existing Description variable.

By removing null values, rejecting variables that add little to no value to analysis, and simplifying/concatenating the text variables available, I felt confident that dataset was prepared to begin model development.

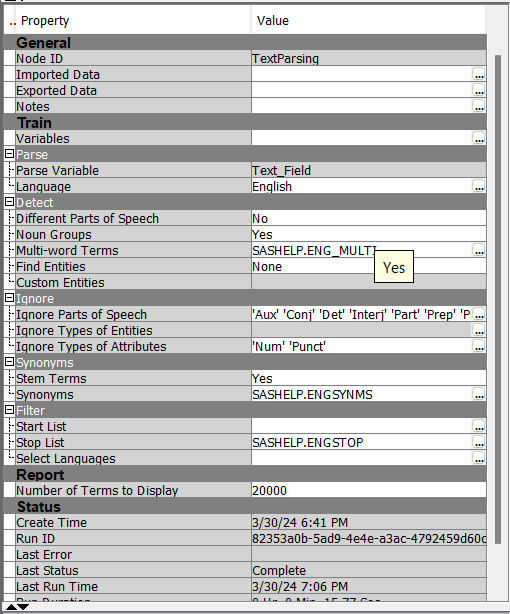
**3.D) Preprocessing the Data**

By leveraging target variable prediction and evaluation workflows explained in class such as the Forensics and Subrogation examples, team aligned on a plan for approaching the nodes leading up to the actual model development.

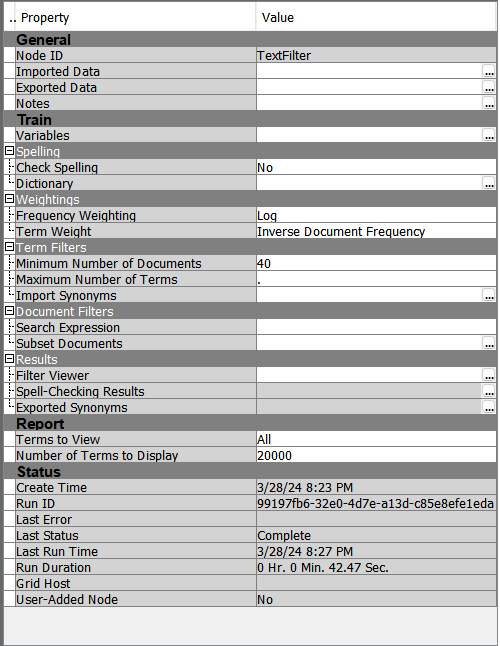
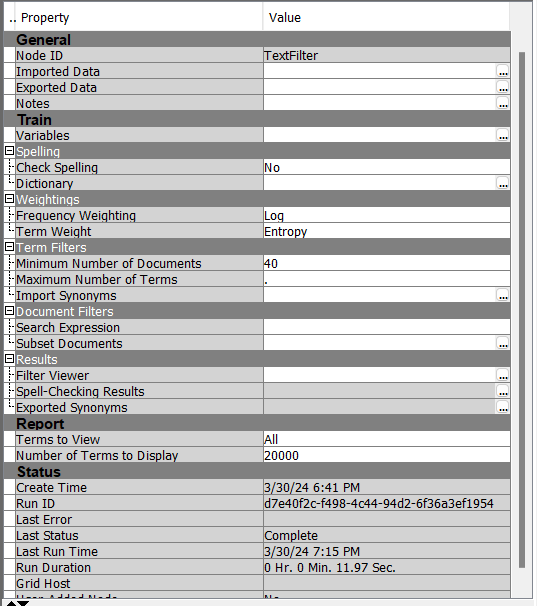
* The **Sample** node alloId for the stratification and creation of random samples from the data, which was crucial for tasks like model training and validation, especially when working with such a large dataset.
* Next, I leveraged the **Data Partition** node to split the data into multiple parts, enabling us to evaluate model performance on testing data. Because data is highly skeId, I selected the Stratified Partitioning Method with Data Set Allocations of 50/30/20 for training, validation, and testing, respectively. This approach helped ensure that each partition maintains the same class distribution as the original dataset, which is crucial for building a predictive model that can accurately mitigate the effects of skeId classes.

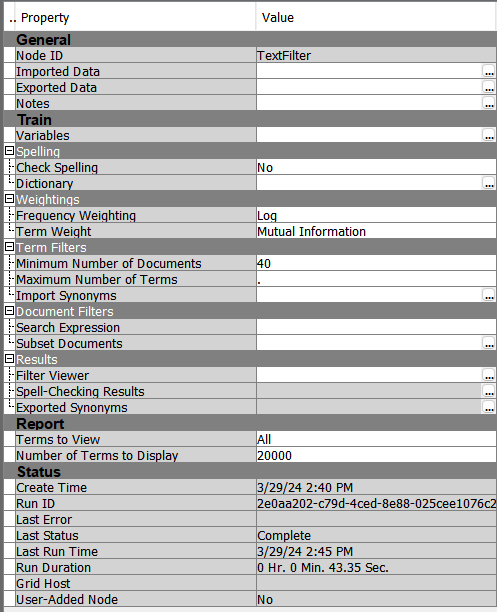


* The **Text Parsing** node was then used to extract meaningful information from text variable, which varied based on the model being evaluated. I opted to use the SAS-provided Multi-word Terms and ignored auxiliary verbs conjunctions, determiners, interjections, articles, prepositions, and pronouns due to their limited value-add ability across the number of documents. I also discovered the SAS-provided Stop Word List was sufficient when trying to eliminate noise across many documents, with minimal changes providing valuable enhancements to model development.



* The **Text Filter** node assisted in removing “noisy” text, ensuring that the analysis focused on the most relevant information. It was during this part of the work when I wanted to pull in several Frequency Iightings and Term Iights to determine which permutation would be most impactful on model accuracy. As a team, I felt comfortable with 40 being the minimum number of documents for each term to appear and evaluated three variations of filters: Log/Mutual Information, Log/IDF and Log/Entropy. goal to include all three in diagram was to allow for a more comprehensive analysis of term importance from different perspectives. Since each measure captures a different aspect of term relevance, including all three branches provides us with a more detailed understanding of the text.
  + Based on class lectures, I wanted Log/Mutual Information to help calculate how much knowing the presence or absence of a term informs whether the job posting is fraudulent.
  + On the other hand, I wanted Log/IDF to help identify rarer terms that could also be informative.
  + And lastly, goal with Log/Entropy was to help highlight terms that occur in many documents but might be less valuable in predicting whether the posting is fraudulent.





**4) MODEL DEVELOPMENT**

team continued model development by attaching nodes that are more aligned with model analysis:

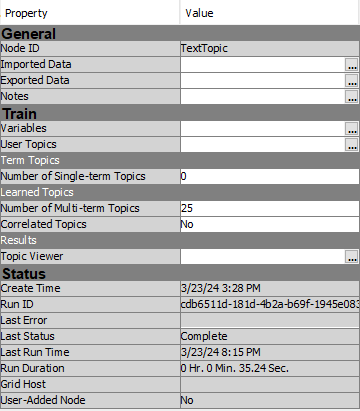
* The **Text Cluster** node was used to group similar text documents together, allowing for the identification of patterns and themes in the data. I evaluated several permutations of SVD Resolution and Max SVD Dimensions to determine the optimal settings for reducing the dimensionality of the text data while preserving essential information for clustering. I know that these selections affect the amount of information retained from the original text data, which could influence decision tree's ability to learn meaningful patterns and make accurate predictions. I explored several variations of SVD Resolutions (Low/High) and Max SVD Dimensions (ranging from 20-120).

A screenshot of a computer

Description automatically generated

|  |  |
| --- | --- |
| **SVD Resolutions** | **Dimensions** |
| Low | 100 |
| High | 20 |
| High | 50 |
| High | 80 |
| High | 120 |

* The **Text Topic** node was used to identify the most prominent topics in a collection of text documents, providing valuable insights into the underlying themes and trends in the data. I opted to use the default settings in preparation for Decision Tree analysis.



Once all the prior modeling nodes Ire attached, team began exploring the types of models I wanted to explore. As discussed in lecture, Decision Trees and Logistic Regressions can be beneficial while evaluating binary target variables. Knowing there are multiple inputs that can affect the output of Decision Trees and Logistic Regression models, I first planned to create an aligned focus on which text field(s) to build the model upon.

Because I previously built the concatenated field, I felt confident to begin evaluating various inputs of Decision Tree and Logistic Regression models. This included Log frequency Iights, term Iights of Entropy/Mutual Information/IDF and a minimum of 40 documents. As discussed previously, I also explored 5 different combinations of SVD Resolutions and the number of dimensions ranging from 20-120. To summarize the 24 different models built:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Frequency Iight** | **Term Iight** | **Min # of Documents** | **Model** | **SVD Resolution** | **# of Dimensions** |
| Log | Entropy | 40 | Decision Tree | Low | 100 |
| Regression | Low | 100 |
| Decision Tree | High | 50 |
| Regression | High | 50 |
| Decision Tree | High | 80 |
| Regression | High | 80 |
| Decision Tree | High | 20 |
| Regression | High | 20 |
| Log | MI | 40 | Decision Tree | Low | 100 |
| Regression | Low | 100 |
| Decision Tree | High | 50 |
| Regression | High | 50 |
| Decision Tree | High | 80 |
| Regression | High | 80 |
| Decision Tree | High | 20 |
| Regression | High | 20 |
| Log | IDF | 40 | Decision Tree | Low | 100 |
| Regression | Low | 100 |
| Decision Tree | High | 50 |
| Regression | High | 50 |
| Decision Tree | High | 80 |
| Regression | High | 80 |
| Decision Tree | High | 20 |
| Regression | High | 20 |

team ran each of the above 24 models thrice, once for each identified text variable:

* Company Description
* Company Profile
* Title + Company Description

**5) MODEL EVALUATION**

**5.A) Performance Metrics**

team evaluated three accuracy measures for each of the models built:

|  |  |  |
| --- | --- | --- |
| **Accuracy Measure** | **Description** | **Calculation** |
| Precision | How many are positive out of everything predicted? | True Positives / (True positives + False Positives) |
| Recall | Out of everything actually positive, how many Ire predicted to be positive? | True Positives/ (True positives + False Negatives) |
| Accuracy | How many accurate predictions Ire made? | True Positives +True Negatives / (True Positives + True Negatives + False Positives + False Negatives)  Or...  1 - Misclassification Rate |

**5.B) Title + Description as Text Field**

When leveraging a concatenated Title + Description field as text variable, I found the highest accuracy measures amongst Log and Entropy’s Low SVD Resolution Decision Tree with 100 SVD dimensions, as Ill as Log and Entropy’s High SVD Resolution Decision Tree with 50 SVD dimensions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | **Model** | **SVD Resolution /Dimensions** | **Recall** | **Precision** | **Accuracy** | **Misclassification rate** |
| Log and Entropy | Decision Tree | Low/100 | 0.66 | 0.81 | 0.80 | 0.19 |
| Regression | Low/100 | 0.01 | 0.75 | 0.60 | 0.39 |
| Decision Tree | High/50 | 0.67 | 0.80 | 0.80 | 0.19 |
| Regression | High/50 | 0.02 | 0.66 | 0.60 | 0.39 |
| Decision Tree | High/80 | 0.68 | 0.78 | 0.79 | 0.20 |
| Regression | High/80 | 0.01 | 1 | 0.60 | 0.39 |
| Decision Tree | High/20 | 0.72 | 0.73 | 0.78 | 0.21 |
| Regression | High/20 | 0.02 | 0.80 | 0.60 | 0.39 |
| Log and MI | Decision Tree | Low/100 | 0.67 | 0.77 | 0.79 | 0.20 |
| Regression | Low/100 | 0.02 | 0.80 | 0.61 | 0.39 |
| Decision Tree | High/50 | 0.67 | 0.77 | 0.79 | 0.21 |
| Regression | High/50 | 0.006 | 0.5 | 0.59 | 0.41 |
| Decision Tree | High/80 | 0.58 | 0.84 | 0.78 | 0.21 |
| Regression | High/80 | 0.12 | 0.66 | 0.60 | 0.39 |
| Decision Tree | High/20 | 0.55 | 0.88 | 0.79 | 0.21 |
| Regression | High/20 | 0.01 | 0.60 | 0.60 | 0.39 |
| Log and IDF | Decision Tree | Low/100 | 0.66 | 0.74 | 0.77 | 0.22 |
| Regression | Low/100 | 0.006 | 1 | 0.60 | 0.39 |
| Decision Tree | High/50 | 0.62 | 0.73 | 0.75 | 0.24 |
| Regression | High/50 | 0.01 | 0.66 | 0.60 | 0.39 |
| Decision Tree | High/80 | 0.61 | 0.73 | 0.75 | 0.24 |
| Regression | High/80 | 0.01 | 0.50 | 0.59 | 0.40 |
| Decision Tree | High/20 | 0.57 | 0.79 | 0.76 | 0.23 |
| Regression | High/20 | 0.01 | 0.60 | 0.60 | 0.39 |

With knowledge that the Log and Entropy’s Decision Tree performed better than other Text Filter and Model combinations across the board, I decided to focus attention on these permutations for the remainder text fields.

**5.C) Company Profile as Text Field**

When leveraging the Company Profile field as text variable, I found the highest accuracy measures amongst Log and Entropy’s High SVD Resolution Decision Tree with 80 SVD dimensions. These metrics are performing better than the concatenated Title + Description explored above.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | **Model** | **SVD**  **Resolution / Dimension** | **Recall** | **Precision** | **Accuracy** | **Misclassification Rate** |
| Log and Entropy | Decision Tree | High/80 | 0.87 | 0.78 | 0.86 | 0.13 |
| Decision Tree | High/100 | 0.87 | 0.78 | 0.85 | 0.14 |
| Decision Tree | High/50 | 0.82 | 0.78 | 0.85 | 0.14 |

**5.D) Description as Text Field**

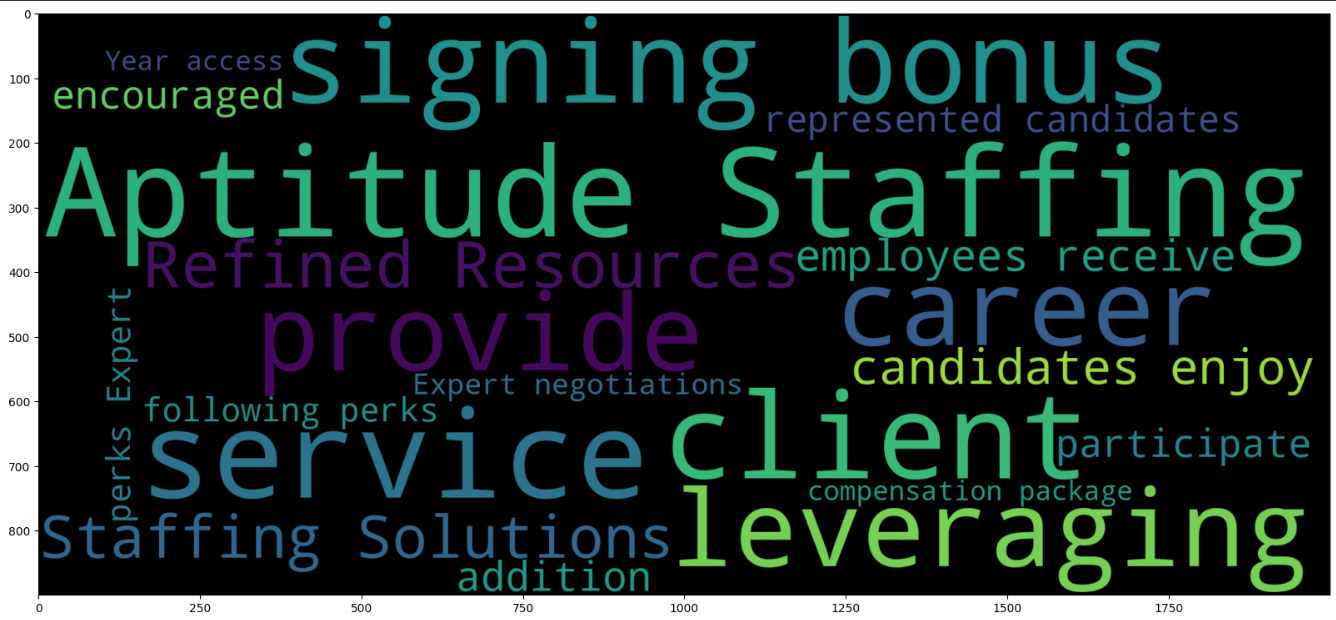
When leveraging the Description field as text variable, I found the highest accuracy measures amongst Log and Entropy’s High SVD Resolution Decision Tree with 120 SVD dimensions. HoIver, although this is the best combination for the Description text field, it still does not perform as Ill as the highest-performing models utilizing the Company Profile field as text variable.

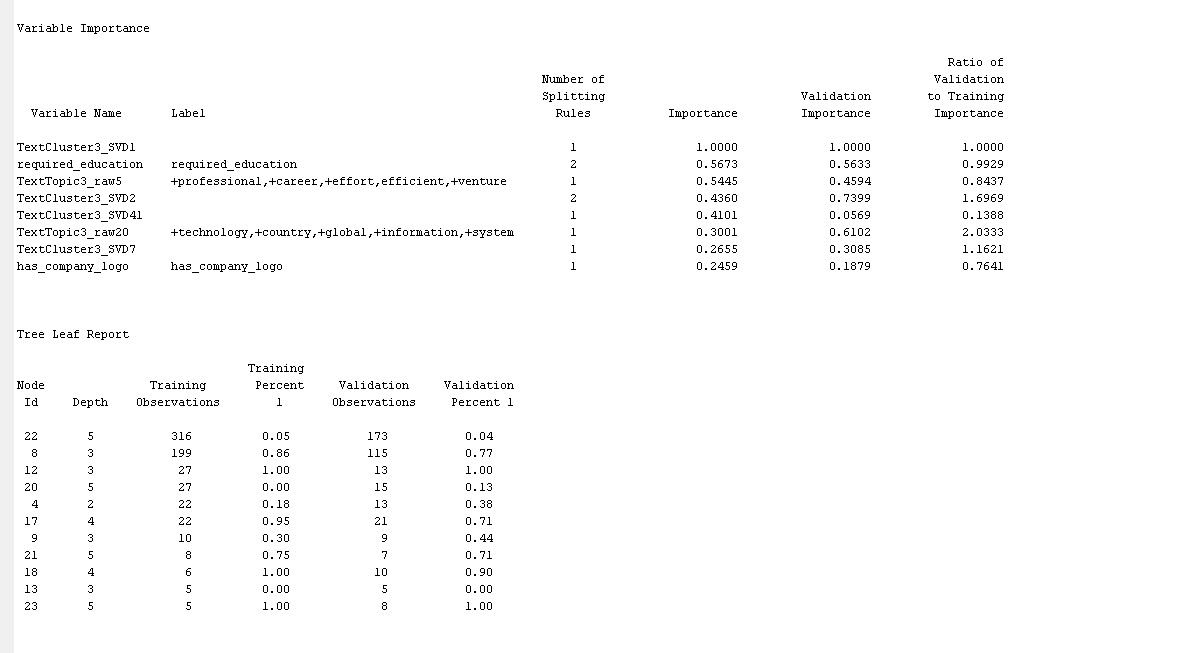
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | **Model Label** | **SVD**  **Resolution / Dimension** | **Recall** | **Precision** | **Accuracy** | **Misclassification Rate** |
| Log and Entropy | Decision Tree | High/120 | 0.76 | 0.78 | 0.82 | 0.17 |
| Decision Tree | High/80 | 0.65 | 0.84 | 0.81 | 0.18 |
| Decision Tree | High/50 | 0.63 | 0.86 | 0.81 | 0.18 |

**6) CONCLUSION & RECOMMENDATIONS**

Out of all the models that I built, I found that the Log and Entropy Decision Tree with high Resolution and a high SVD Dimension of 120 produces the best accuracy of 82% and a misclassification rate of 17%. The recall rate of 76% suggests that this model correctly identifies 76% of the actual fraudulent postings whereas precision emphasizes 78% of jobs labeled as fraudulent Ire found to be fraudulent.

By performing an analysis in SAS as Ill as a word-cloud in Python, I Ire able to detect familiar words being used to make fraudulent job postings more attractive and appealing:





The common phrase "signing bonus" suggests that these jobs may attract potential candidates with the promise of immediate financial rewards. Additionally, the mention of "perks," "compensation package," and "enjoy" suggests that these jobs promise numerous benefits to entice job seekers. Words like “staffing,” "refined," "expert," "resces," and "service" are used to create an illusion of professionalism and legitimacy. More common words such as "career" and "aptitude" are used to appeal to individuals looking for personal and professional development.

**6.A) Recommendations for Job Seekers**

As best model suggests, job seekers should pay special attention to company profile above anything else. Job seekers should watch out for the cautionary words outlined above. An example of a company profile that may be fraudulent based on this recommendation includes:

*innovative new platform cuts the recruiting time in half, yields scientifically proven results and clients and candidates* ***enjoy*** *a pleasant experience through advanced, simple to use technology and a tenured, industry-experienced recruiting team. Join us in a fresh new experience of leveraging y* ***career****...the way it should be! All represented candidates* ***enjoy*** *the following* *perk****s****:* ***Expert*** *negotiations, maximizing total* ***compensation package****.* ***Signing bonus*** *by Aptitude Staffing in addition to client* ***signing bonus*** *(if applicable)1 Year access to AnyPerkRelocation Services for out-of-town candidates.*

If a job posting seems suspicious, the candidate should contact a contact for validation prior to sharing any personally identifiable information. On a surface-level analysis, I also found that most fraud job companies do not have a company logo.

**6.B) Recommendations for Job Sites**

Some general recommendations for job sites are they should enroll in Quality Control Teams to assess the job postings thoroughly before posting them on their Website. These teams may monitor new job postings for the above caution words in company profiles to proactively flag potential job postings for users. If a potentially fraudulent company has multiple job postings, they may choose to enact a suspension period is appropriate while the team evaluates each individual job posting for relevancy and validity.

**7) REFERENCES**

Fair, Lesley. “Job Scammers Go Even LoIr in the Way They “Hire.”” *Federal Trade Commission*, 3 Aug. 2023, [www.ftc.gov/business-guidance/blog/2023/08/job-scammers-go-even-loIr-way-they-hire?tpcc=NL\_Marketing](http://www.ftc.gov/business-guidance/blog/2023/08/job-scammers-go-even-lower-way-they-hire?tpcc=NL_Marketing).

Gurchiek, Kathy . “Beware: Fraudulent Job Postings Aim to Steal Y Money, Identity.” [*Www.shrm.org*](http://Www.shrm.org), 10 Aug. 2023, [www.shrm.org/topics-tools/news/beware-fraudulent-job-postings-aim-to-steal-money-identity#:~:text=Fraudsters%20posing%20as%20recruiters%20and%20employers%20are%20contacting](http://www.shrm.org/topics-tools/news/beware-fraudulent-job-postings-aim-to-steal-money-identity#:~:text=Fraudsters%20posing%20as%20recruiters%20and%20employers%20are%20contacting).

Smith, Sean M., et al. “Unemployment Rises in 2020, as the Country Battles the COVID-19 Pandemic : Monthly Labor Review: U.S. Bureau of Labor Statistics.” *U.S. Bureau of Labor Statistics*, June 2021, www.bls.gov/opub/mlr/2021/article/unemployment-rises-in-2020-as-the-country-battles-the-covid-19-pandemic.htm.