Capstone 2: Final Project Report

Problem Statement:

Which skills should I focus on learning first, where should I focus my job search, and in which industry should I focus my job search in to increase my odds of getting a job as a Data Scientist or Data Analyst in 2024?

Context:

Data Scientist and Data Analyst jobs require a specialized set of skills. I want to identify the most in demand skills in the Glassdoor job descriptions for a Data Scientist and Data Analyst, so I know which ones to focus on developing first to be a more highly rated candidate during my job search. Understanding where these jobs are located and in which industries will allow me to more efficiently focus my search in the appropriate state and industry to increase my odds of getting a job as a Data Scientist or Data Analyst.

Data:

Kaggle is an online community of data scientists and machine learning engineers. Kaggle allows users to find datasets they want to use in building AI models, publish datasets, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. The data for this dataset was collected by RRK-CODER (Kaggle profile name) through a systematic web scraping process using the Selenium framework. Leveraging Selenium allowed for the automated retrieval and analysis of information from over 1,500 live data science job listings on Glassdoor.com.

The data encompasses essential details about each job listing, facilitating comprehensive analysis and insights into the data science job market. I decided to focus the analysis on the skills, location, industry, and company ratings for each open position to allow me to prioritize which skills to learn first, which industry is hiring the most data scientists/data analysts, and where these positions are located in the United States (including remote positions).

Glassdoor Data Science Job Listings

Method:

There are many types of classification algorithms out there in the wild, but I decided to start with 3 supervised classification models.

- Logistic Regression: Logistic Regression is a simple yet effective linear classifier that works well for text classification tasks. It's interpretable, allowing you to determine which skills have the highest weights, which can indicate their demand. It is also computationally

efficient and provides easily interpretable results that highlight the most significant features/skills.

- Random Forest Classifier: Random Forest is a powerful ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. It can also give feature importance scores, which can be used to identify which skills are most relevant across job descriptions. Random Forest models are capable of handling a large amount of data and are less sensitive to overfitting compared to individual decision trees.

- Stacked Model (LightGBM + XGBoost): LightGBM (Light Gradient Boosting Machine) and XGBoost are both ensemble methods based on boosting, which means they create a series of weak learners (typically decision trees) that focus on improving errors from previous iterations. This makes them effective at handling complex patterns in data. Since LightGBM and XGBoost may learn different aspects of the data, stacking helps create a balanced final model as well as mitigating the risk of overfitting by reducing the variance.

I chose to work with the Random Forest model (after trying all 3) in the end because it is a strong middle ground, offering good performance for a wide variety of tasks while being more scalable and efficient than the stacked model. It balances complexity and predictive power, making it a good general-purpose model, especially when computational resources are available.

Data Cleaning:
To get the data into a better state for analysis and predictive modeling, first I needed to filter out any jobs that did not include "Data Scientist" or "Data Analyst" in the job title. Although some jobs with different titles could essentially be data scientists or data analysts based on the job descriptions, I wanted this to be explicit and not leave any additional room for interpretation regarding job title. I was mostly focused on the skills needed, and this filtering still resulted in over 1,100 records of job titles with Data Scientist or Data Analyst which I deemed sufficient for this exercise.

Next I needed to identify the locations of the job openings by state with "Remote" also being considered a location. This was accomplished by a user-defined function that stripped away the city and replaced it with the 2 letter abbreviation for the state. Then a new column "State" was created and updated with the 2 letter state abbreviation which also included "Remote".

Now a list of skills needed to be defined, so I could make predictions on which ones to focus on learning first. While this is a subjective selection, I still feel it is a good selection of skills and software to know based on my previous 8 months of researching open positions. I chose what I considered to be the top 20 most common ones I have seen as well as the ones that appeared in this dataset. In fairness, a more exhaustive list could be created by counting all of the skills

and software listed in the dataset. I would also consider updating this list on a yearly basis to capture any new skills or emerging technology/software to stay current.

EDA:

- There were 1166 job titles containing Data Scientist and 32 containing Data analyst.
- Remote jobs lead the way with 218, but California was close with 195 job openings which makes up ~16% of the job openings for Data Scientist and Data Analyst positions.
- I.T. Support Services and Health Care are the leading industries hiring for Data Scientists and Data Analysts and made up 22% of the openings out of the 68 unique industries hiring.
- The pay range for both positions was typically between $100,000 to $200,000.
- The company ratings didn't seem to vary by state or industry and maintained a median score of 3.9.

Modeling:

After testing and tuning all 3 classification models (Logistic Regression, Random Forest, Stacked LightGBM + XGBoost), I chose the Random Forest model for its scalability, predictive power, and efficiency (especially compared to the stacked model).

The tuned Random Forest model with class balancing generally shows improvements in recall and F1 scores, particularly for difficult-to-predict classes. The accuracy remains stable overall. However, precision dropped for a few cases, possibly due to the class balancing leading to more false positives. The model performed better for skills like Target_Spark, Target_NLP, and Target_Data Mining, but it still struggles with certain skills like Target_Azure, Target_Keras, and Target_Power BI.

Conclusion:

- The Accuracy values are generally high for most skills, with many above 0.90, which indicates that the model correctly predicts many of the outcomes. However, accuracy alone can be misleading, especially with imbalanced datasets, as it may not reflect how well the model performs on minority classes.
- Precision and Recall scores show significant variability across skills, suggesting that while the model is good at predicting the presence of some skills, it struggles with others.

Summary

The analysis indicates variability in the model's ability to predict different skills accurately. Some skills are well predicted, while others are not detected at all. Improvements can be made by focusing on better feature representation, addressing class imbalance, optimizing model

parameters, and utilizing more advanced NLP models. Emphasis should also be given to enriching the dataset to ensure that all skills are sufficiently represented.

Future Work:

1. Further Data Preprocessing and Feature Representation:
   - Feature Representation: Improve the representation of text data by exploring advanced natural language processing (NLP) techniques such as TF-IDF or Word embeddings (e.g., Word2Vec, BERT), which may help capture the context in which skills are mentioned better than simpler techniques.
   - Balancing the Dataset: Although class balancing was attempted, it may need further refinement. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN can be used to better handle imbalanced classes for skills that show poor performance.
2. Model Optimization:
   - Hyperparameter Tuning: The Random Forest model may benefit from additional hyperparameter tuning. Consider optimizing parameters such as the number of trees (n_estimators), depth of trees (max_depth), and the number of features (max_features) to help improve recall and precision for underperforming classes.

Recommendations:

1. Use the model to identify the most in demand skills in the job market for Data Scientist and Data Analysts by determining which skills are the most important.
2. This model can be used to help curate a training curriculum based on the most important skills to upskill individuals looking to work as a Data Scientist or Data Analyst.
3. Help understand which industries are hiring the most Data Scientists and Data Analysts and how the companies are rated (according to GlassDoor) to aid job seekers in focusing their job search.