ISYE 6740 – Summer 2022
Final Report

Member: Mariam Ammar

Project Title: Analysis of data jobs success factors

Problem Statement

Due to the increase in demand and relatively higher salaries for data-related roles, many fresh graduates and students are seeking to enter the field of data. However, given the abundant amount of tools and variety of data-related roles, it can be intimidating for newcomers to focus on key characteristics that would contribute to their success within the field and determine how much their skills are worth in the end.

Although countless articles have been written by professionals on which tools and areas of focus are deemed crucial for success in the field, there is limited updated, quantitative analysis available for the general public that answers two questions in which this project explores:

1. What are the factors that contribute to higher salaries within data-related roles (gender, years of experience, knowledge of specific languages/tools, etc.)
2. Given all the characteristics that an individual possesses, what salary should they receive?

Previous Research

This dataset has been used for a variety of explorations and analysis such as correlations between salary and happiness and gender discrepancies, however none of the notebooks featured using this data focus on the specified aforementioned problem statements with a focus on the United States job market. This notebook as well as this notebook conduct analysis of factors that contribute to higher salaries within the data science and other related fields. However, they do not include all features that are

included in this project and do not focus on one country and therefore skew the salary earnings since salaries in India and countries outside of the US tend to vary significantly.

Data

Each year, Kaggle releases a “Machine Learning and Data Science Survey” in an attempt to create a comprehensive view of the current situation in the data science and machine learning fields by surveying individuals from around the world. The survey was live from 09/01/2021 to 10/04/2021 and received 25,973 responses. The questionnaire consists of 38 questions.

Methodology

Data Wrangling

This project uses a filtered dataset consisting of 13 features that correspond to answers to 13 questions asked in the questionnaire that are relevant to the questions being asked. These answers or features are also simplified to allow for ease of interpretation. For example, answers that feature several intervals were transferred to feature lesser intervals of a broader range. In addition, questions that feature several answer choices were simplified by selecting the top n choices and creating an ‘other’ category. For a detailed description of the features to be included, please see Appendix A.

The target variable ‘salary’ was simplified to seven classes to mitigate data imbalance issues. In addition, the SMOTE(Synthetic Minority Oversampling Technique) was used. This technique works by utilizing a KNN algorithm to create synthetic data for minority classes. It works by generating instances that are close in feature space, using interpolation between positive cases that are close to each other and randomly selecting the minority class instance to find its nearest neighbor. Once the technique was implemented, there were 375 data points within each of the 8 classes within the training dataset.

Modeling

The following classification models were chosen: Logistic Regression, KNN, Kernel SVM, Naive Bayes, RandomForest, and Adaboost. The metric of accuracy score was used for model selection in this case since in this context false positives and false negatives are both weighted equally. In addition, each model was cross-validated.

Model	Accuracy	Accuray with Cross-Validation
Logistic Regression	38.53%	39.55%
KNN	32.68%	33.19%
Kernel SVM	32.68%	41.22%
Naive Bayes Classifier	37.23%	37.96%
Adaboost	32.03%	32.68%
Random Forest	35.93%	38.87%

From all the previous model runs, the Kernel SVM model achieved the highest accuracy. However, feature importance of linear SVMs can be found out but not for a nonlinear SVMs, the reason being that, when the SVM is non-linear the dataset is mapped into a space of higher dimension, which is quite different from the parent dataset and the hyperplane is obtained and this high dimensional data and hence the property is changed from that of the parent dataset and hence it is not possible to find the feature importance of this SVM in relation to the parent dataset features.

The second highest scoring model in terms of accuracy was the Logistic Regression model. However, the stats module in Python for logistic regression only creates output for binary classification problems. Due to these reasons and the fact that accuracy for the Random Forest model does not show significant deviation (less than 2%), this model was chosen for ease of explainability to move forward with the hyperparameter tuning and variable selection.

Hyperparameter Tuning

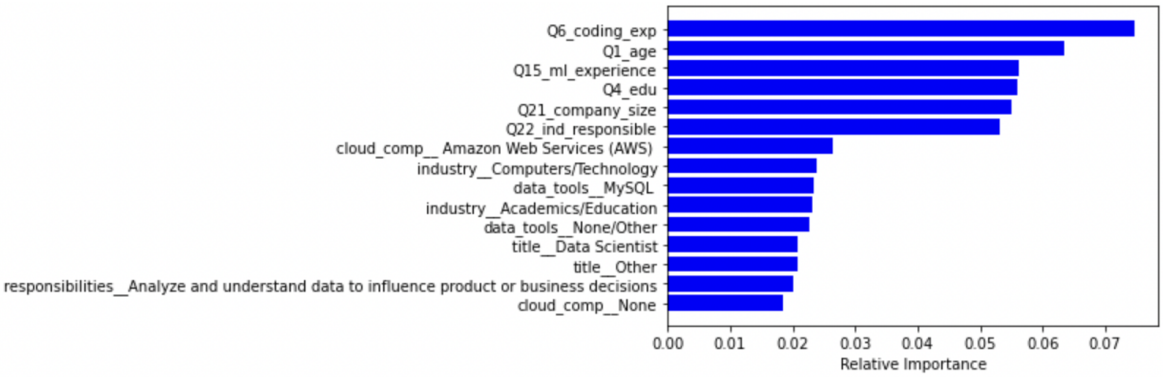
After utilizing the best model from a Randomized Search CV, accuracy was increased from 38.87% to 42.86% using the following parameters.

{'n_estimators': 1800,
 'min_samples_split': 2,
 'min_samples_leaf': 4,
 'max_features': 'sqrt',
 'max_depth': 90,
 'bootstrap': True}

Randomized Search CV implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter values. This is unlike GridSearchCV where every combination of a preset list of values of the hyper-parameters is tried and the model is evaluated for each combination. For this project, the Randomized Search CV was the better performer.

Feature Importance

While it is important to note the low accuracy, the analysis suggests that six main features are most impactful when predicting salary since these features came out to be the most impactful features from the featured Random Forest model and one and one can note significant drop off for the following feature importances.



The screenshot above suggests that years of coding experience, age, years of machine learning experience, education, company size, and the number of individuals

responsible for data-related roles within a company may impact the salary that one receives within the data field.

Decision trees perform feature selection implicitly by computing the decrease in entropy or impurity from one note to the next. Nodes towards the top of the tree are weighed more than those at the bottom. An average of all the differences in impurities is then calculated and normalized so that they equal 1. Basically by doing this one is able to pinpoint the features that cause the greatest amount of variation (decrease the entropy/impurity) within the data.

While certain features such as those that include experience and age are likely correlated, the Random Forest model is not significantly affected by multicollinearity since it chooses different sets of features for different trees and these estimators see different sets of data points. However, correlated features share importance making it difficult to note particularly which feature (in this case machine learning experience or education) has a greater impact in predicting salaries since either feature can be used to reduce the same amount of impurity for classification. Another caveat that should be noted of this method considers the combination of several one-hot-encoded variables and ordinally encoded features. Since one-hot-encoded features can only be used once within each tree, this has the potential to place higher importance on ordinal or continuous variables vs. the former.

Evaluation and Final Results

Overall, this analysis suggests that the years of coding experience, age, years of machine learning experience, education, company size, and the number of individuals responsible for data-related roles within a company have the highest impact on the salary that one receives within the data field.

One should also note a variety of other factors which include the low accuracy generated by this model, any noise that may result from self-input answers from survey takers, bias that may result from the use of several one-hot-encoded features,

correlation between features that may skew feature importance, and the consideration of features that were not included in this project.

For validation, it is recommended to acquire data from a similar survey that features a continuous target variable with exact salary amounts or to acquire objectively input data by combining salary information from years of experience and education scraped from sources such as LinkedIn. In addition, one can re-consider combining intervals of inputs such as was done in this project to gain a more granular analysis. Lastly, it is also important to consider that each case that determines salary may not be codependent on specific features. For example, someone with higher levels of education could earn the same salary as an individual without the same level of education but with more years of experience.

References

https://www.kaggle.com/competitions/kaggle-survey-2021/overview

https://www.kaggle.com/code/toysperfect/beginners-guide-in-there-data-science-journey

https://www.kaggle.com/code/muhammadgusanwaakbar/formula-to-be-successfull-in-data-science

https://www.kaggle.com/code/janalvin/are-you-happy-correlation-of-happiness-and-salary

https://www.kaggle.com/code/rloredo/gender-equality-exploration

Appendix A

Feature	Question #	Description	Comment
Age	1	Range	Simplified ranges
Gender	2	Man Woman Nonbinary Prefer not to say Prefer to self-describe	Top 2 and other
Education	4	Highest Level of Education	Top three or other
Title	5	Title most similar to your current role	Top 5 and other
Experience	6	Years writing code/programming	Simplified Range
Languages	7	Programming languages used as a regular basis	Top 3 and other
ML_experience	15	Years of experience using ML models	Simplified ranges
Industry	20	Industry of current employer/contract	Top 5 and other
Company Size	21	Size of the company currently employed in	
Individuals Responsible	22	Individuals are responsible for data science workloads at place of business	Simplified Range
Important responsibilities	24	Analyze and understand data to influence product or business decisions Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data Build prototypes to explore applying machine learning to new areas Build and/or run a machine learning service that operationally improves my product or workflows Experimentation and iteration to improve existing ML models Do research that advances the state of the art of machine learning	Top four and other

		None of these activities are an important part of my role at work	
Cloud Computing Platforms	27-A	Cloud computing platforms used on a regular basis	Top 3 and other
Big Data Products	32-A	Big data products (relational databases, data warehouses, data lakes, or similar) used on a regular basis	Top 5 and other
Salary	25	Current yearly compensation in USD	Simplified into 8 intervals

The full dataset can be downloaded via this link.