Ronald Pacheco

Final Report:
Glassdoor Salary Prediction

Problem Statement

Executive Summary

In this project, we are looking to create a supervised regression model that can predict the salary for a Data Science position. This is useful to compare current compensation with the market, or to determine where the best opportunities are.

Background

This project will involve missing predictive values and missing salary in job posting by the companies, which are the constraints for identifying the best city, title, keywords to predict salary and correlation between variables.
Principal stakeholders in this project are Ronald Pacheco and Blake Arensdorf.

Success Criteria

This project will be successful once a model that predicts salary is created, along with a project report.

Data Sources:

- https://www.kaggle.com/andrewmvd/data-analyst-jobs
- https://www.kaggle.com/andrewmvd/business-analyst-jobs
- https://www.kaggle.com/atharvap329/glassdoor-data-science-job-data?select=Data_Job_SF.csv
- https://www.kaggle.com/andrewmvd/data-scientist-jobs

Data Wrangling

The data consisted of 13,585 observations distributed amongst 7 datasets and two different formats (variables named differently). First step was to consolidate them all in one dataset, this required the variables to be renamed.

In the process, data cleaning feature engineering were applied to create better features, such as splitting the target variable into 'Min_Salary' and 'Max_Salary' (later converted to 'Avg_Salary'), location was split into 'State' and 'City', the job description was parsed to extract keywords like 'Python', 'SQL', 'R', etc.

Missing variables were imputed using KNNImputer, and categorical variables were one-hot encoded. This resulted in a dataset of shape (11440, 140).

Exploratory Data Analysis

The relationship between the dependent variable and independent variables were studied, but no patterns were found:

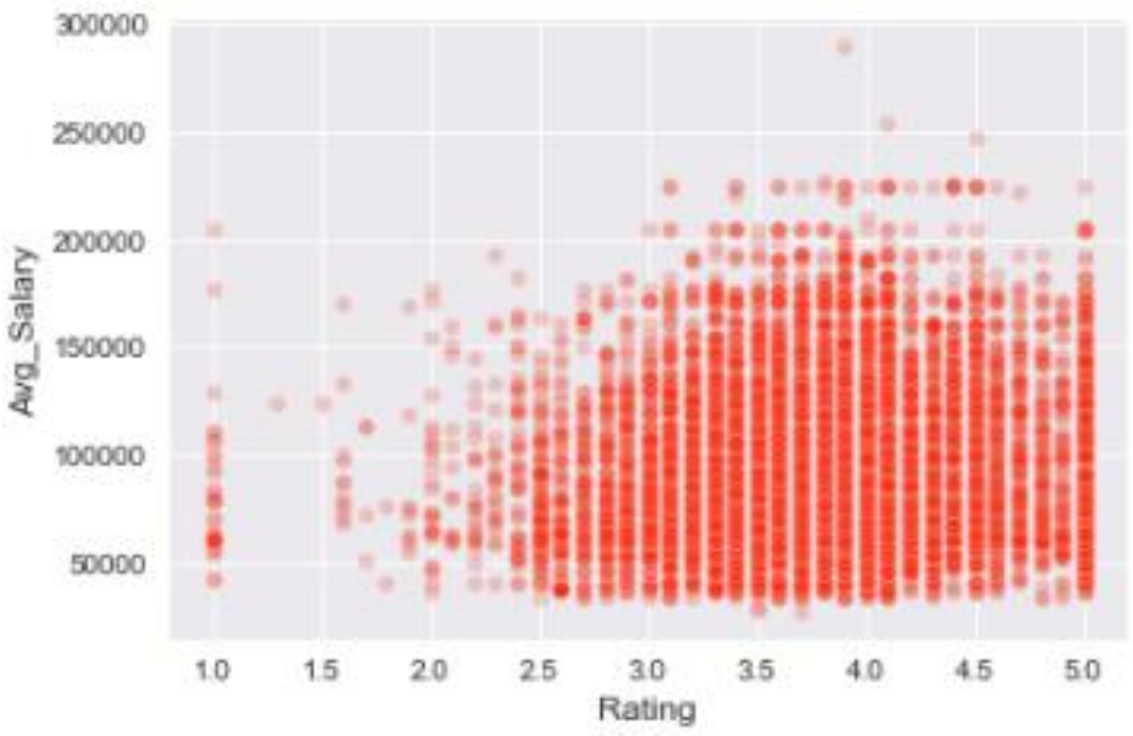

Figure 1. Rating vs Avg_Salary

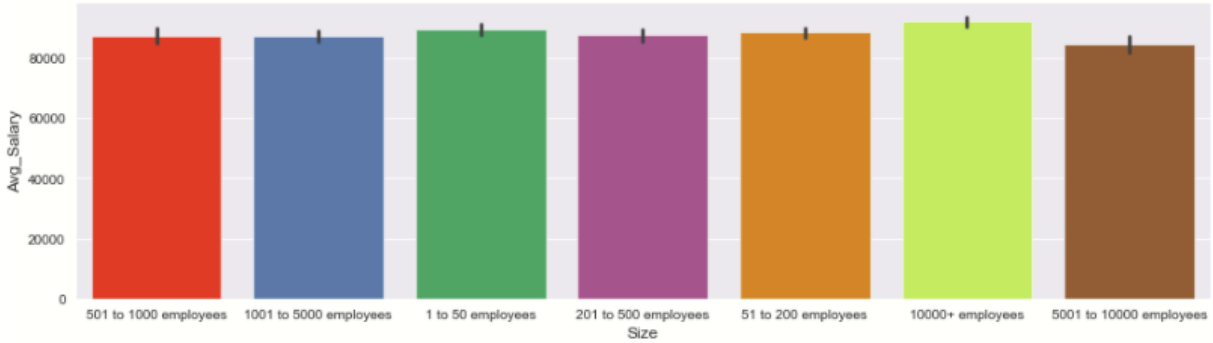

Figure 2. Size vs Avg_Salary

However, we were able to get insights from our data:

- The mean average salary is $91K, the max is $290K and the min is $27.5K.
- Rating, Size & Revenue do not seem to affect the average salary for the position.
- Sectors Biotech & Pharmaceuticals, Agriculture & Forestry & Media seem to pay more on average.
- On average, the 5 highest paying states are: CA, KY, VA, NY and MD. Contrarywise, UT, GA, IN, KS and WA seem to be the worst paying states.
- GA is not representative of the population.
- Julia and Python seem to have the most positive correlation with average salary.
- SQL and Scala seem to not make a significant difference.
- JavaScript is actually negatively correlated with average salary.
- California is hiring the most Data Scientist that know Python.
- Biotech & Pharmaceuticals, Agriculture & Forestry & Media seem to pay more on average.

Model Selection

For this regression project, the models tested were RandomForestRegressor, Lasso, XGBoost, and LinearRegression. The metric used to select the best performing model was MAE (Mean Absolute Error).

RandomForestRegressor

Our initial model, with no hyperparameter tuning, gave a MAE of $21,492. After hyperparameter tuning, the MAE was $20,941.

The model is: RandomForestRegressor(n_jobs=-1, random_state=42, bootstrap= True, max_depth= 10, max_features= 'auto', min_samples_leaf= 5, n_estimators= 1000).

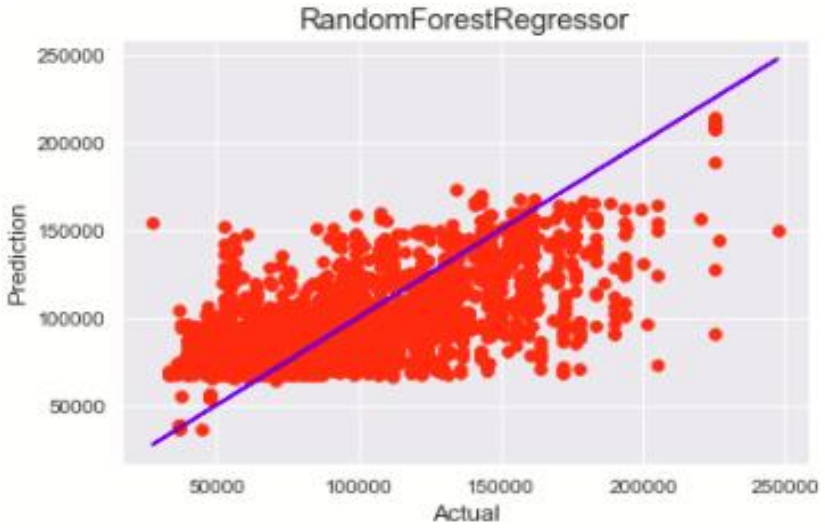

Figure 3. RandomForestRegressor Prediction vs Actual

Lasso

Our initial model, with no hyperparameter tuning, gave a MAE of $21,744. After hyperparameter tuning, the MAE was $21,739.

The model is: Lasso(alpha=1).

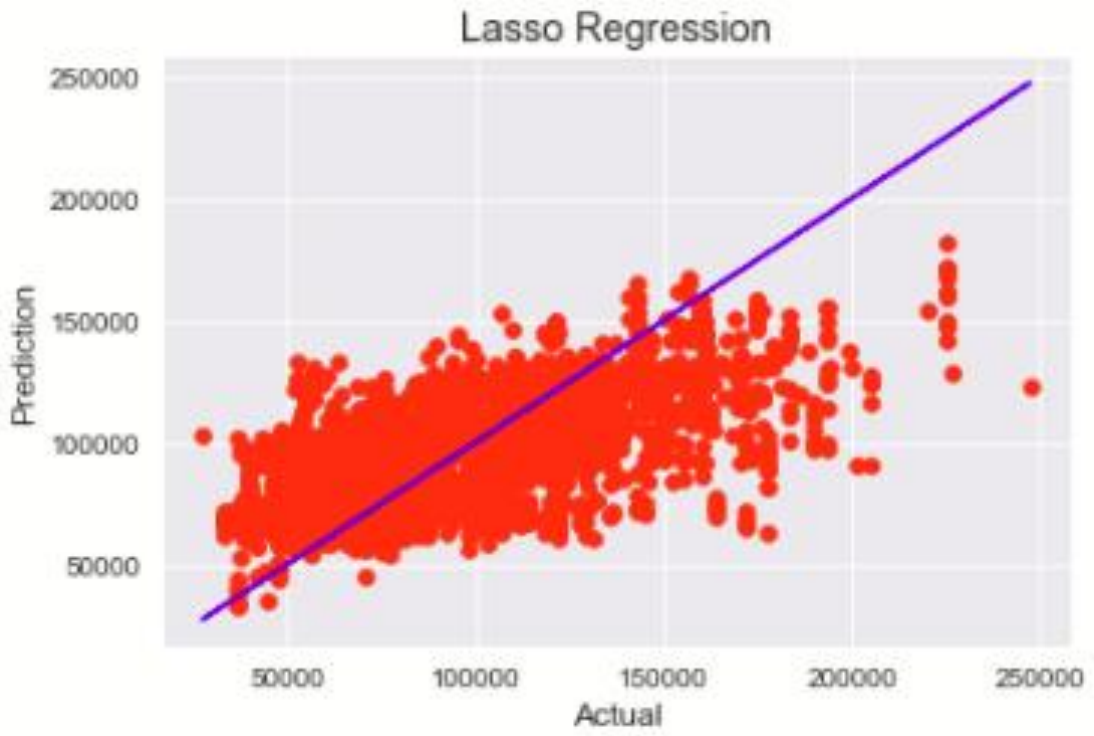

Figure 4. Lasso Prediction vs Actual

XGBoost

Our initial model, with no hyperparameter tuning, gave a MAE of $21,238. After hyperparameter tuning, the MAE was $21,048.

The model is: XGBRegressor(learning_rate=0.01, max_depth=10, n_estimators=1000, objective='reg:squarederror', seed=123).

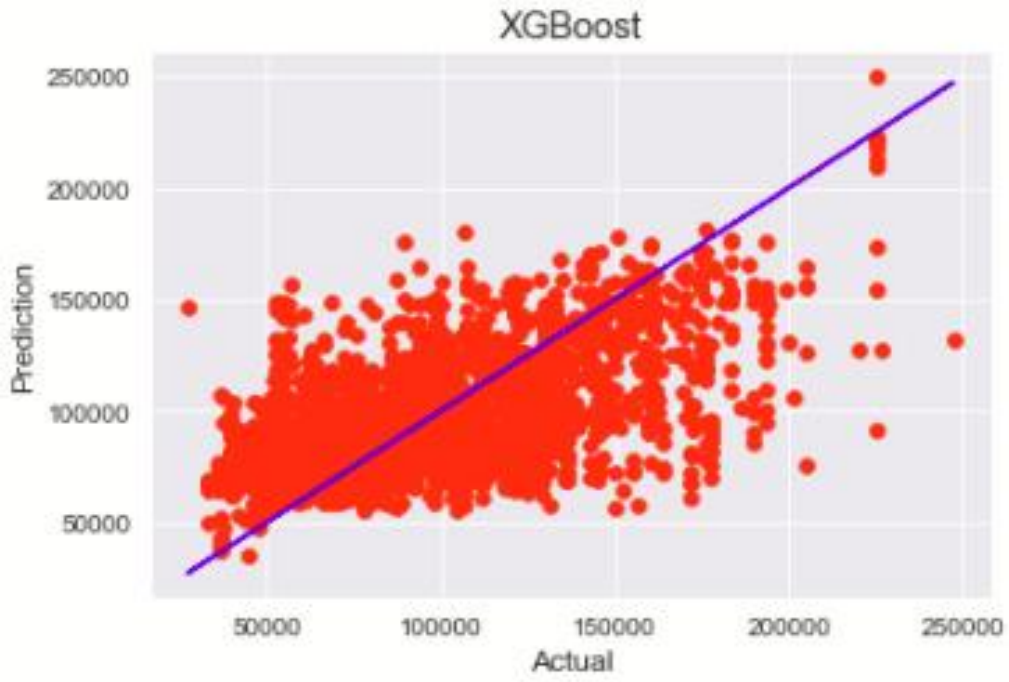

Figure 5. XGBoost Prediction vs Actual

LinearRegression

Our initial model, with no hyperparameter tuning, gave a MAE of $21,749. After hyperparameter tuning, the MAE was $21,696.

The model is: Pipeline(steps=[('selectkbest', SelectKBest(k=79)), ('linearregression', LinearRegression())]).

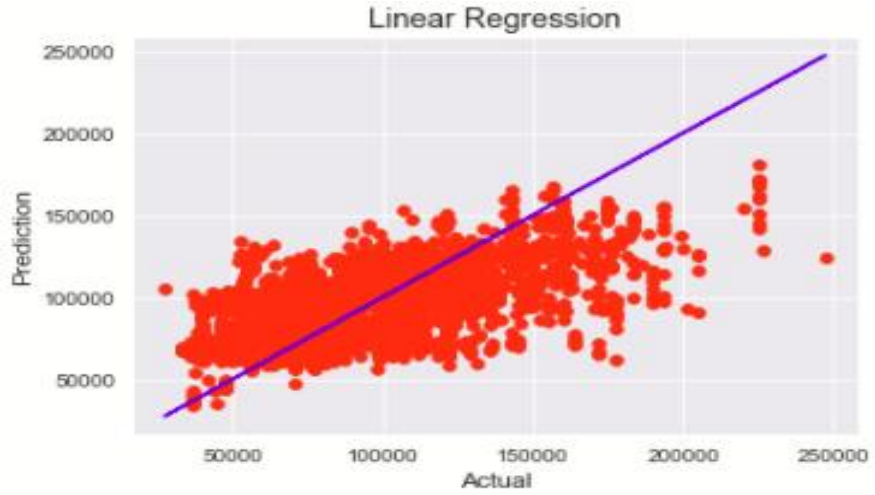

Figure 6. LinearRegression Prediction vs Actual

Conclusions

- Our best model ended up being a RandomForestRegressor.
- Our MAE is $20941. I believe this is due to our dataset. Most observation had a Glassdoor Estimate range with a difference greater than $40K, and, despite this, our model is able to reduce that range to slightly below $42K.

Additional approach to take

- NLP can be applied to Job_Desc to try and find correlations with the actual text and the salary. We tried a version of this by extracting keywords such as Python, R, Scala, etc.