Data Science Job Analysis

Overview

As Data Science field is exponentially growing; Number of job seeker has spiked up and so, there is a need to closely analyze the job market to know about different job sectors of Data Science, industry, job location, job title and its salary range.

Problem Statement:

Data Science Job analysis project is to evaluate various aspect of Data Science skill market such as

- Find best jobs by Salary, Company Rating and Location.
- Chances of getting hired in different job sectors (data scientist, data engineer and data analyst)
- Skills needed to get hired in a specific job sectors and salary range.
- Which company hires the most Data Science professionals?
- Best hot spot (cities) of all.

Dataset

The datasets link is mentioned below:

https://www.kaggle.com/andrewmvd/data-scientist-jobs

- For this project I obtain the data from Kaggle. This dataset was containing data science positions (all assumed to be open positions at the time the dataset was published in July 2020), with features such as: Salary Estimate, Location, Company Rating, Headquarter and more.
- Dataset has lots of missing value in a form of " -1", "unknown". I have taken care of those missing values and replace it with most appropriate values.

Methodologies

To understand how the variation in an independent variable can impact the dependent variable, regression analysis is used.

1. First perform Exploratory Data Analysis (distribution, boxplot, bar charts, heatmaps, scatterplots...etc.) to get relationship between variables.
2. Multiple Regression As this is a regression problem, I build multiple regression model to evaluate which one gives us better result.
3. Dataset containing lots of categorical data as independent variables. Create dummy variables for each value whose presence in the dataset.
4. After splitting training/ testing datasets, perform multiple regression.
5. Compare the models and make the final decision. After comparing all the model Random Forest is winner.

Data Cleaning

There are a couple of things I would like to clean up with this data set to lend itself to more thorough analysis:

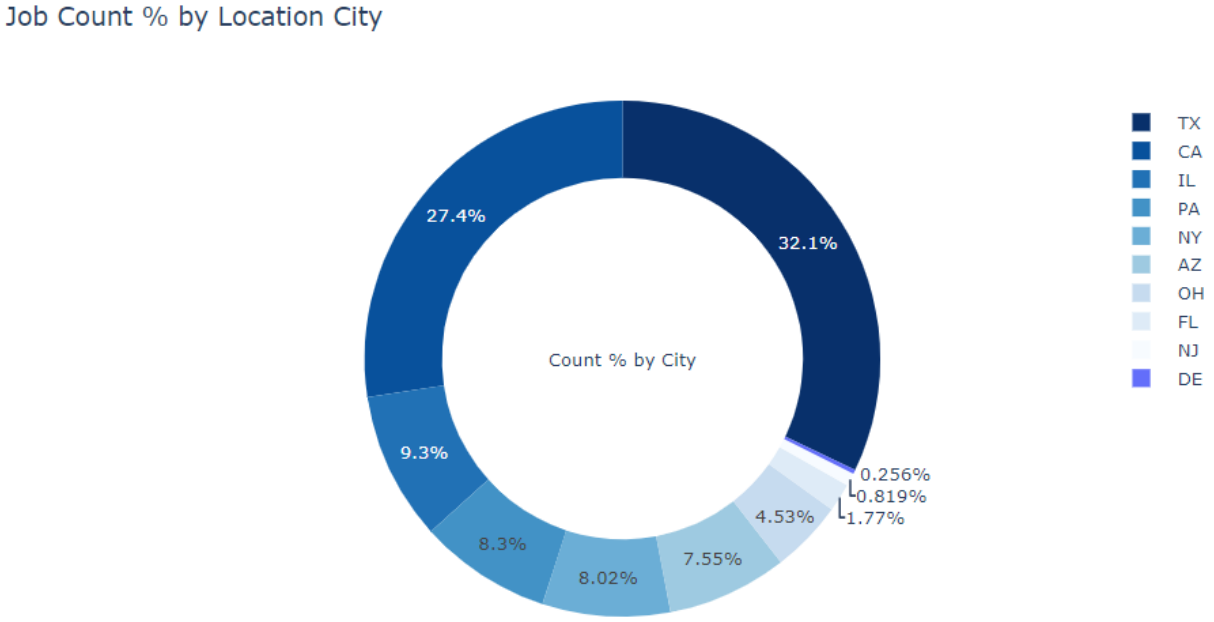
- Remove the Unnamed: 0 Column
- Split Salary Estimate into salary_estimate_upper_bound, salary_estimate_lower_bound, and extrapolate per hour to annual salary.
- Clean Company Name Split the "\n & rating" from company Name. (remove what appears to be the rating)
- Split Size into Lower Bound and Upper Bound
- Create 3 major categories (Data Scientist, Data Engineer, Data Analyst) of data science job role among all.
- Replace all -1 & 0 into more appropriate data. For example, missing value in rating column is replaced with its mean value. In Sector column “-1” value is replaced with unknown sector.

EDA (Exploratory Data Analysis)

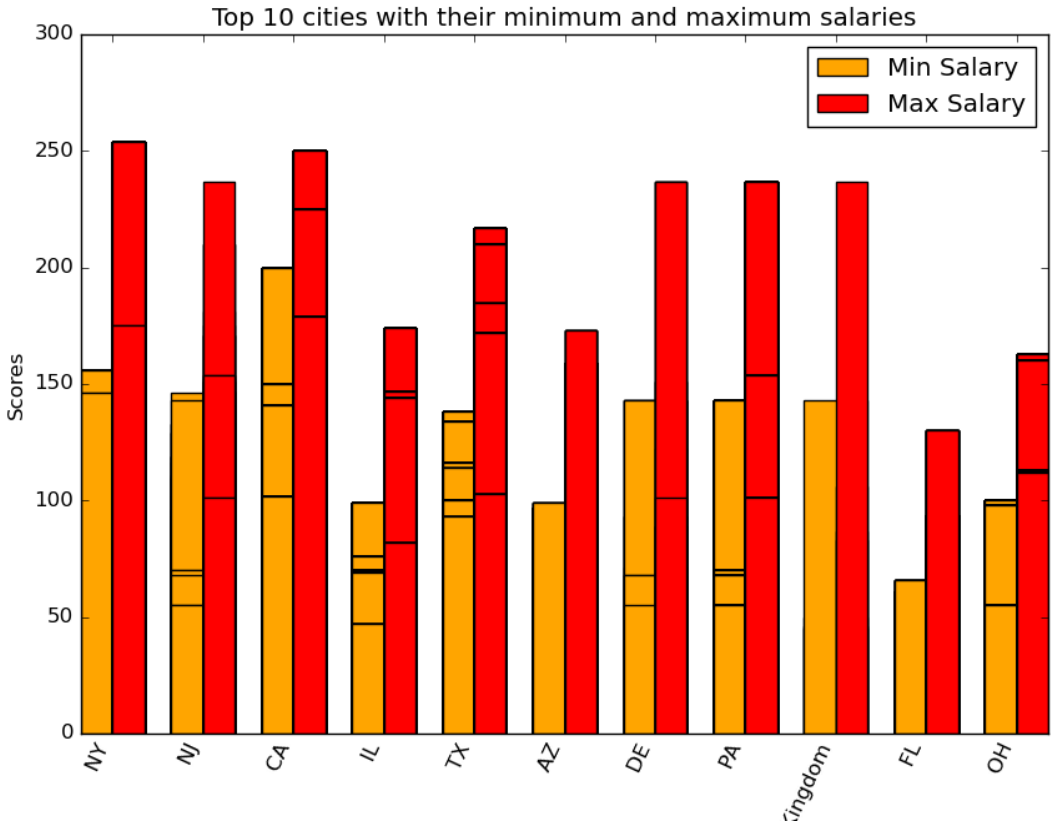
In this step I visualize the dataset and did lots of finding like

Name of the Company Involved in Data Science, most Popular Data Science Job Title, Top 20 locations for Data Science Job, Top 20 Head Quarters of Data Science, Job Holder Company. Number of companies founded in year, Types of ownership of company, Different sector involve in data science job, Job count by city location.

After analyzing the top 10 state Texas have most job, California comes in second position. Most popular data science job title is Data Scientist, Data engineer and data analyst. Most of the company is owned by private sector.



Top 10 state maximum and minimum salary.



Preprocessing

80% data (3127 out of 3678) was reserving for training and 20% (782 Out of 3678) data is reserved for testing. We often need to prepare our data in specific ways before feeding into a machine learning model, we must convert all the categorical independent variables into dummy variable for modeling task.

Using get_dummies: - The get_dummies method transforms categorical data into binary columns. Applied get_dummied to 14 columns is converted to 3678 dummy columns.
After applying get dummies I have 3678 columns our first task is to do feature selection. When I perform feature, selection get to know salary is more dependent on location and job title.

Algorithms & Machine Learning

I chose to work with the Python Regression Algorithm because in regression problem output variable is a real or continuous value. I tested 4 different regression.

1. Linear Regression
   Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).
   Linear regression test accuracy is
2. Ridge Regression

3. Lasso Regression

4. Random Forest

Modelling Result

Regresion	MAE	MAPE	RMSE	Test data accuracy
Linear	44.07	25.31	41.73	-1.17
Ridge	26.65	28.91	32.48	27.75
Lasso	28.46	31.01	33.95	21.06
Random Forest	24.35	25.31	31.75	30.95

NOTE: I choose RMSE as the accuracy metric over mean absolute error (MAE) because the errors are squared before they are averaged which gives the RMSE a higher weight to large errors. Thus, the RMSE is useful when large errors are undesirable. The smaller the RMSE, the more accurate the prediction because the RMSE takes the square root of the residual errors of the line of best fit.

Also check the accuracy more the value more accurate the model is.

Winner: Random Forest

Conclusion

Location State_ CA
Location State_ NY
Location City_San Diego
Founded
Location City_Los Angeles
Job Title_Data Scientist
Job Title_Data Analyst
Job Title_Data Engineer
Location City_Austin

After feature importance get to know Salary is mostly dependent on location after that role come in picture. If anyone looking data science role in CALIFORNIA salary must be higher than NY. Salary also depends on how establish company is means how old is the company. Data Scientist salary must be higher than Analyst and engineer position.

Recommendation for client

According to our best model our salary prediction shows

- if our predicted value is very **high** then we subtract **"Negative test residual"** i.e. 56 to get the right **prediction**.
- if our predicted value is very **low** then we add **"Positive test residual"** i.e. 70 to get the right **prediction**.

Limitations and Assumptions

- The results only reflect the outcome at the time the dataset was published, which is presumed to be July 2020. Seasonal variation is disregarded (not a time-series data).
- The salary estimates come from Glassdoor, which may not reflect the actual salaries.
- The dataset is assumed to reflect the traits of the actual job market.