**INTRODUCTION**

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. It is a branch of Artificial Intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

A Data Model is built automatically and further trained to make real-time predictions. This is where the machine learning algorithms are used in the Data Science Lifecycle.

Scope of Data Science:

The influence of data science is enormous in multiple sectors. Every below-mentioned has a high dependence on the data science and offers great opportunities for a data scientist.

Sectors recruiting data scientists are:

Healthcare, Marketing and advertising, Airlines, Automobile industry, Virtual reality and augmented reality, IT etc.

To help Data Scientists in negotiating their income when they get a job, I have made a tool that estimates the salary for Data Scientist.

For the project I have scraped approximately 1000 jobs data from ‘www.glassdoor.com’ through Web scrapping using python and selenium. Web scraping is the process of using bots to extract content and data from a website.

I have engineered features from the text of each job description to quantify the value companies put on python, excel, AWS, and spark.

Also I have optimized the Linear regression model, Lasso regression model, and Random forest regressor model using GridsearchCV to reach the best model. I have used Python version 3.9.7 in this project.

**PROBLEM STATEMENT AND OBJECTIVE**

1. **Problem Statement:**

To predict Data Scientist’s salary based on different parameters like type of post, company location, company ratings, company size, company revenue, no. of competitors, company age, skills required, seniority etc.

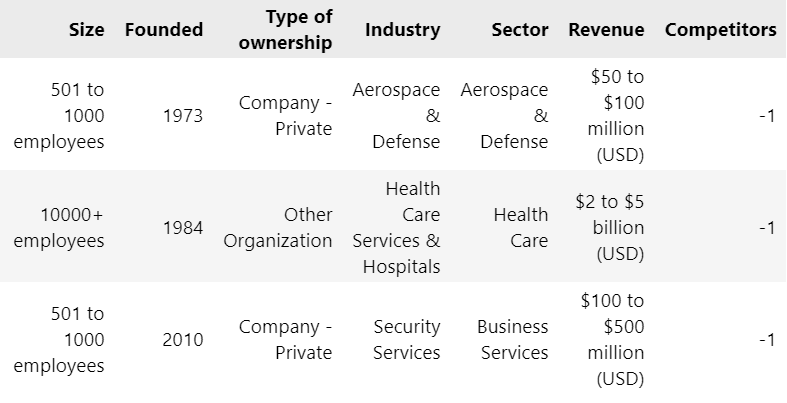
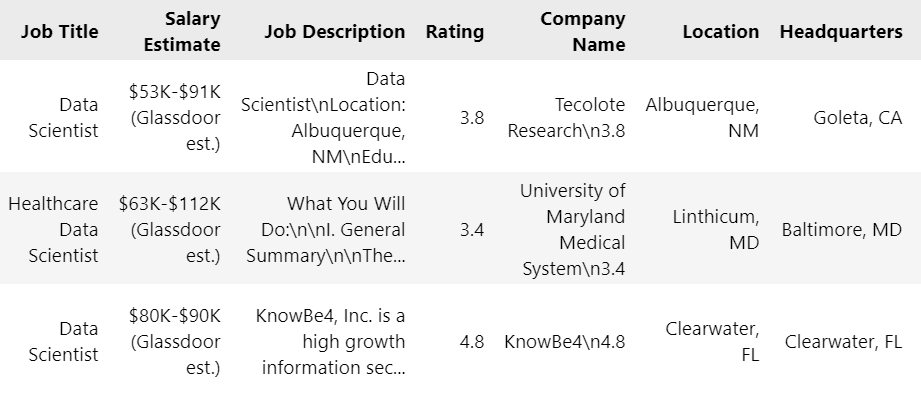
1. **Objectives:**

* To extract data from a job posting site.
* Analysing the data
* Cleaning the data
* Exploratory Data Analysis
* Data Pre-processing
* Model Training
* Optimizing the Model
* Evaluating and Comparing Models

**DESCRIPTION OF THE DATASET**

I have used web scraper to extract the data of approximately 1000 jobs from ‘www.glassdoor.com’. The raw data which I got has following:

* Job title
* Salary Estimate
* Job Description
* Rating
* Company
* Location
* Company Headquarter
* Company size
* Company Founded year
* Type of ownership
* Industry
* Sector
* Revenue
* Competitors



Total no. of attributes: 14

Total no. of entries: 956

Categorical Features:

Job title, Job Description, Company Name, Location, Headquarters, Type of ownership, Industry, Sector.

Numerical Features:

Salary Estimate, Rating, size, Founded, Revenue, competitors.

Where ‘Salary Estimate’ is the target feature.

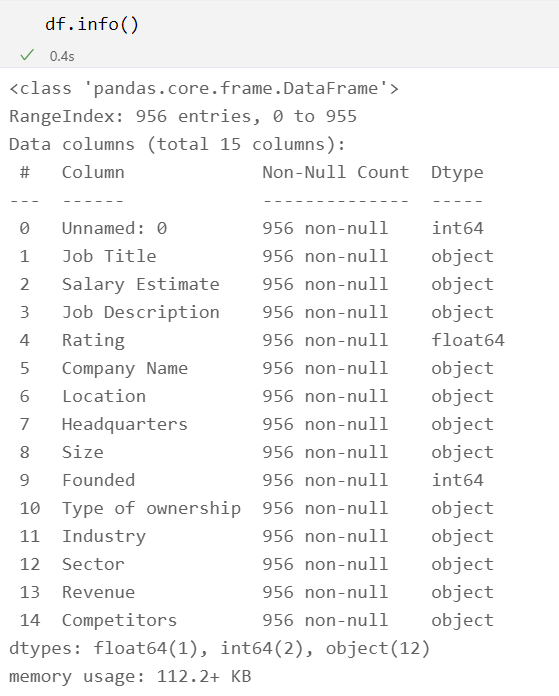
**APPROACH**

I have followed these steps to solve the problem:

STEP 1: Web scraping:

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique used to automatically extract large amounts of data from websites and save it to a file or database. The data scraped will usually be in tabular or spreadsheet format.

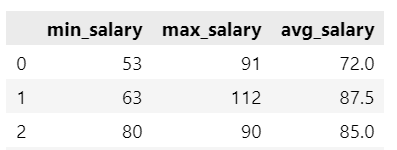
STEP 2: Exploring the web-scraped dataset:



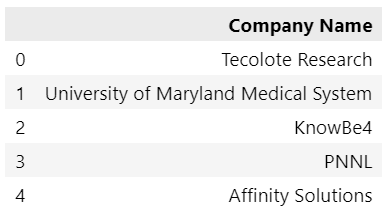
STEP 3: Data cleaning:

Data cleaning refers to identifying and correcting errors in the dataset that may negatively impact a predictive model. Data cleaning is used to refer to all kinds of tasks and activities to detect and repair errors in the data.

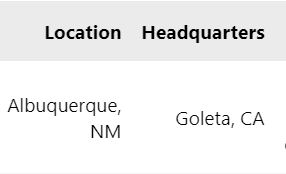
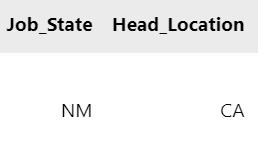
* Salary Parsing: Divided salary range into min salary and max salary and also calculated average salary.

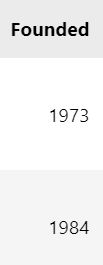
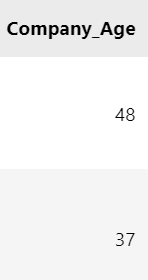
* Company name test only:

* State field: I have separated state from location and headquarter.

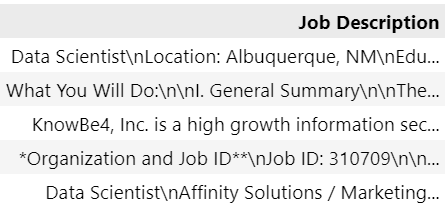
 

* Age of the company: Calculated company age from founded year of the same.

* Parsing of Job Description:

I have extracted information regarding skills (Python, AWS, R studio etc.) required to get job.





STEP 4: Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patters, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

STEP 5: Data Pre-processing:

* Choosing relevant features.
* Encoding categorical variables into numerical variables.

STEP 6: Model Training:

As the target feature i.e. 'Salary' is a continuous variable. Hence I have performed regression analysis on the dataset. I have used 3 machine learning model. They are the follwings:

* Multiple Linear Regression Model:

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

* Lasso Regression Model:

The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

* Random Forest Regression Model:

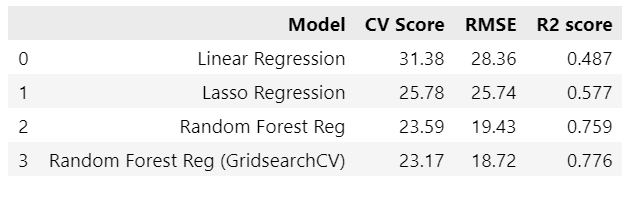
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

STEP 7: Model Optimization:

I have performed ‘Hyperparameter tuning ‘ of Random forest regressor model using GridsearchCV. Grid search is the process of performing hyperparameter tuning in order to determine the optimal values for a given model.

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

STEP 8: Model evaluation and comparison:



Where,

CV Score:

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. CV score is calculate on the basis of Root Mean Square error, hence **lower CV score indicates better performance.**

RMSE:

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. **Lower RMSE indicates better performance.**

R2 score:

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. **The greater r-square value indicates the high correlation between the input variables and output variable and data is suitable for regression model.**

Conclusion:

Clearly Random Forest Regressor model with hypertuning outperformed other models (i.e. Multiple Linear Regression, Lasso Regression)

**Note:** I have also explained all the steps in Jupyter-Notebook which I am going to include.