SALARY PREDICTOR

Data Science Final Project

MEMBERS

Maham Tariq	17K-3658
Eisha Tir Raazia	17K-3730

INTRODUCTION

A salary predictor is created using the job listings from an employment website, in this case Glassdoor.com. A data mining technique is used to generate a model which will scrape the number of jobs from the employment website, clean it on the basis of a number of factors including the rival companies, revenue and skill required thereby predicting the salary to be expected when applying for a data science job. Techniques like linear regression, lasso regression, random forest regressors are optimised using GridsearchCV to reach the best model. The model can be further extended to build a flask API thus can be deployed on the internet for public usage.

METHODOLOGY AND STEPS INVOLVED

1. Setting the research goal
2. Retrieving Data
3. Data Preparation
4. Data Exploration
5. Data Modeling
6. Presentation and Evaluation

1. Setting the research goal

The research goal was to create a tool that estimates data science salaries (MAE ~ $ 11K) to help data scientists negotiate their income when they get a job.

2. Retrieving Data

We scraped over 1000 job descriptions from glassdoor using python and selenium. It took almost 5 days to completely run the code of scraping the data. With each job, we got the following:

- Job title
- Salary Estimate
- Job Description
- Rating
- Company
- Location
- Company Size
- Company Founded Date
- Type of Ownership
- Industry
- Sector
- Revenue

Data Preparation

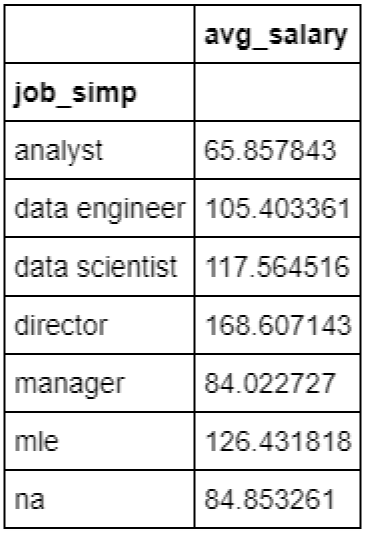
After scraping the data, it was needed to be cleaned so that it was usable for the model. We made the following changes and created the following variables:

- Parsed numeric data out of salary
- Made columns for employer provided salary and hourly wages
- Removed rows without salary
- Made a new column for company state
- Transformed founded date into age of company
- Made columns for if different skills were listed in the job description:
  - Python
  - R
  - Excel
  - AWS
  - Spark

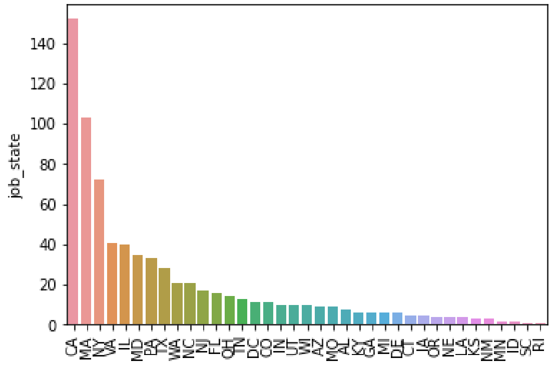
- Column for simplified job title and Seniority
- Column for description length

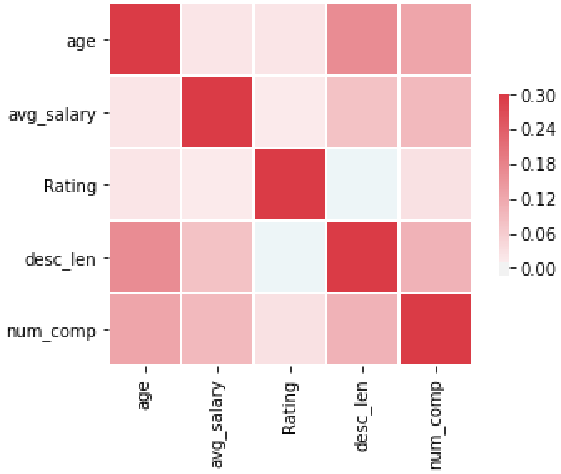
Data Exploration

We looked at the distributions of the data and the value counts for the various categorical variables. Below are a few highlights from the pivot tables.



	avg_salary
job_simp	
analyst	65.857843
data engineer	105.403361
data scientist	117.564516
director	168.607143
manager	84.022727
mle	126.431818
na	84.853261





Data Modeling

Model Building

First, we transformed the categorical variables into dummy variables. We also split the data into train and test sets with a test size of 20%.

We tried three different models and evaluated them using Mean Absolute Error. I chose MAE because it is relatively easy to interpret and outliers aren't particularly bad for this type of model.

We tried three different models:

- Multiple Linear Regression – Baseline for the model
- Lasso Regression – Because of the sparse data from the many categorical variables, I thought a normalized regression like lasso would be effective.
- Random Forest – Again, with the sparsity associated with the data, thought that this would be a good fit.

Presentation and Evaluation

Model performance

The Random Forest model far outperformed the other approaches on the test and validation sets.

1. Random Forest : MAE = 11.22
2. Linear Regression: MAE = 18.86
3. Ridge Regression: MAE = 19.67

DATA

LOREM IPSUM	DOLOR SIT	
Lorem ipsum		
Lorem ipsum		
Lorem ipsum		

RESULTS

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius.

1. Lorem ipsum dolor sit amet
2. Consectetuer adipiscing elit
3. Sed diam nonummy nibh euismod

CONCLUSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

REFERENCES

1. Lorem ipsum dolor sit amet
2. Consectetuer adipiscing elit
3. Sed diam nonummy nibh euismod