Samit Singh

Salary Prediction based on Job Description

Agenda

Salary Prediction

Background / Business Problem

The current job Market is complex. For a HR who wants to hire for certain position & for a candidate who is looking for job change, knowing the salary in advance will be helpful.

Situation	▪ HR wants to hire for certain position in company and for this he/she wants to know the salary in advance. It will help the company to maintain budget. Similarly, a candidate can also estimate his/her current salary or can target certain company or position if he/she can estimate the salary in advance.
Complication	▪ As dataset is huge and there are different factors on which salary depends like education of person, number of years of experience in the industry, designation in the company, type of industry, how many miles a person lives away from city etc. So, Memory optimization will be a challenge along with performance of model.

Executive Summary / Key Takeaways

Approach & Solution

The given dataset has 7 features on which salary depends i.e. Job type, degree, major, industry, company id, years of experience and miles from metropolis.

- We verified which feature has highest impact on the Salary using Pearson correlation and initially, we took the mean of salary by grouping five features (Job type, degree, major, industry & company id) and considered it as baseline.
- The job type is positively correlated with salary, whereas Miles from Metropolis is negatively correlated with salary i.e salary decreases with increase in distance from city.
- There are no duplicates & missing values in the data. But some outliers are present, we need to further analyze the data and remove the outliers.

As there are multiple features which are correlated with salary, we should perform feature engineering to create new meaningful features out of these.

Data Set Characteristics

Dataset Information

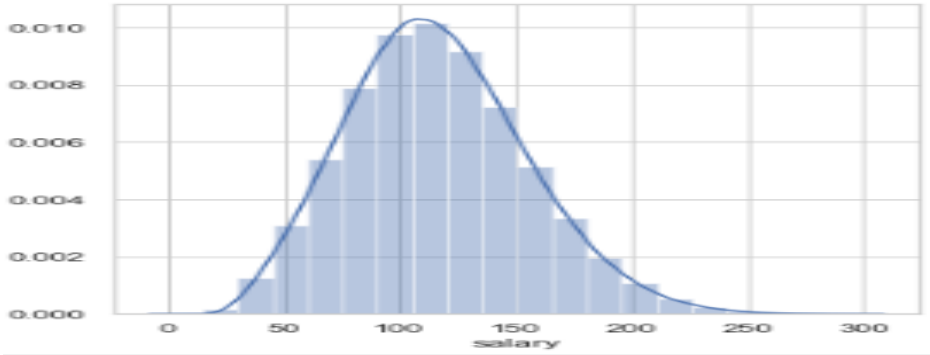
The dataset has 1000000 unique rows & 9 columns.
Following are the characteristics of feature variables:

- jobId – It is unique for every row in dataframe.
- companyId – 63 different companies are present.
- jobType – Following are 8 different categories*.
  CEO > CTO, CFO > VP > Manager > Senior > Junior > Janitor
- Degree: Following are 5 different categories*.
  Doctoral > Masters > Bachelors > High_School > None
- Major: Following are 9 categories*.
  Engineering > Business > Math > Compsci > Physics > Chemistry > Biology > Literature > None
- Industry: Following are 7 categories*.
  Oil > Finance > Web > Health > Auto > Service > Education
- YearsofExperience: Values ranges from 0 to 24.
- milefromMetropolis: Values ranges from 0 to 99.
- salary: After removing zero salary values, it ranges from 17 to 301 thousands dollar per year.

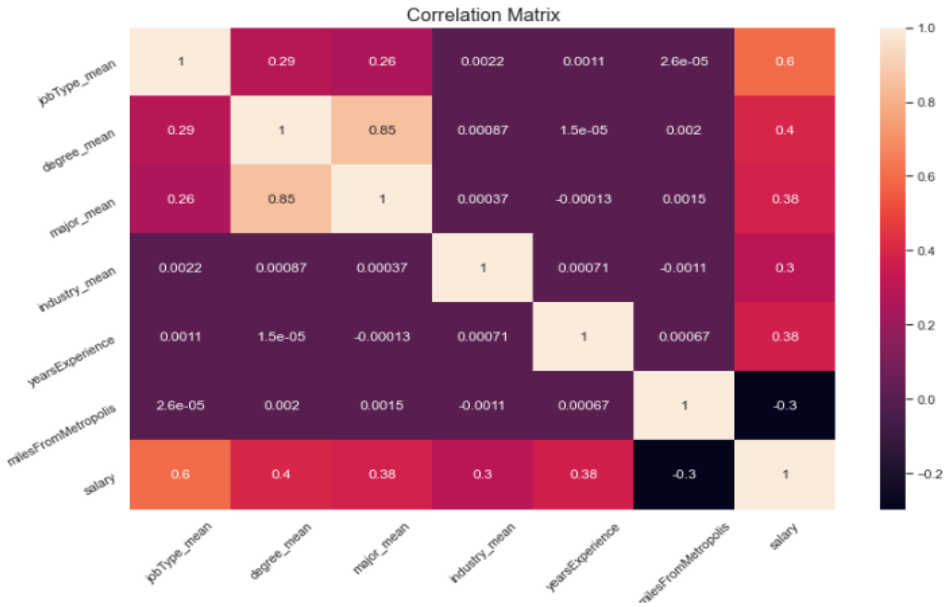
* The categories are sorted in decreasing order with salary.

Dataset Visualizations

- Salary distribution



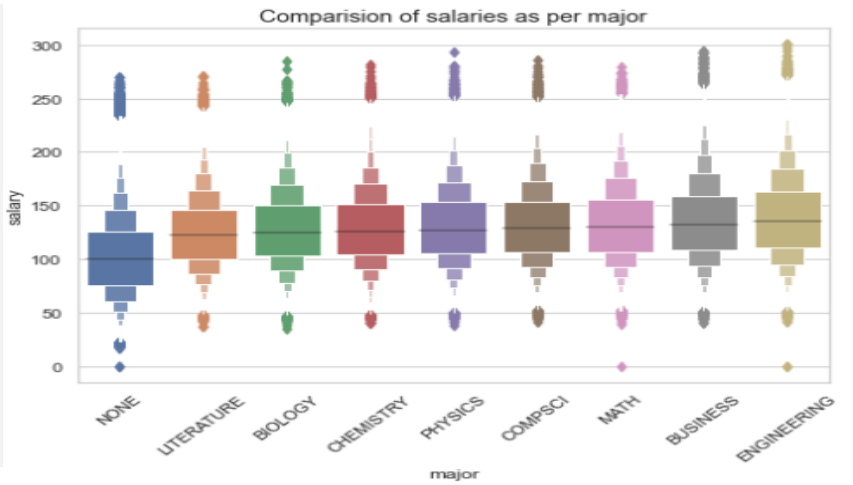
- Correlation Matrix



EDA – Exploratory Data Analysis

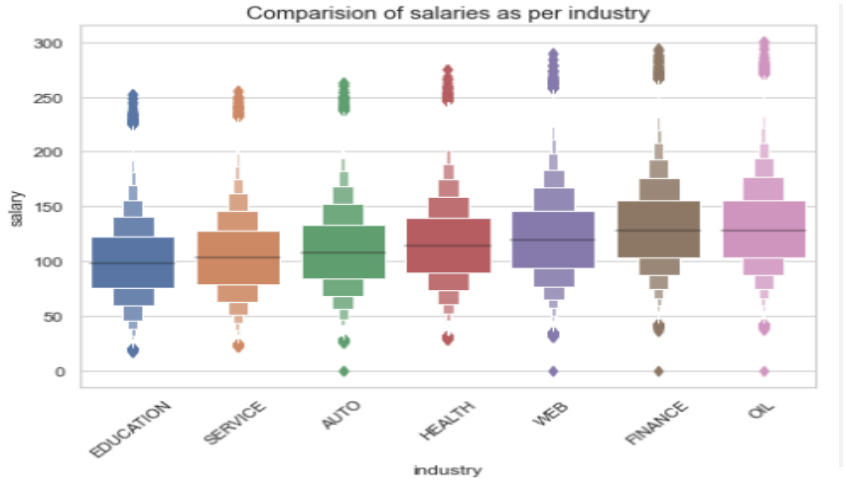
Comparision of salary - JobType

Comparison of salary - Major



Comparision of Salary - Degree

Comparision of salary - Industry



Data Cleansing & Pre-processing

Categorical Features

- ‘jobType’, ‘degree’, ‘major’, ‘industry’ & ’companyId’.
- Label Encoder is used to convert categorical data to numeric.

companyId	jobType	degree	major	industry
COMP37	CFO	MASTERS	MATH	HEALTH
COMP19	CEO	HIGH_SCHOOL	NONE	WEB
COMP52	VICE_PRESIDENT	DOCTORAL	PHYSICS	HEALTH
COMP38	MANAGER	DOCTORAL	CHEMISTRY	AUTO
COMP7	VICE_PRESIDENT	BACHELORS	PHYSICS	FINANCE

Numerical Features

- Years of Experience, Miles from Metropolis & Salary
- 5 rows having 0 salary were removed from dataset.

yearsExperience	milesFromMetropolis	salary
10	83	130
3	73	101
10	38	137
8	17	142
8	16	163

Feature Engineering

- Grouped all Categorical features and computed grouped statistics like ‘average salary’, ‘median salary’, ‘minimum salary’, ‘maximum salary’ & ‘mean absolute deviation salary’ of the group.
- Missing values are replaced by taking group of features except ‘companyId’, as it is least correlated to salary

group_mean	group_median	group_max	group_min	group_mad
130.875000	136.0	150	98	14.906250
138.031250	132.0	204	90	24.378906
142.000000	137.0	173	116	20.666667
129.000000	131.0	169	77	20.571429
153.583333	141.5	232	94	38.680556

Modelling, Tuning & Evaluation

Model Selection

- As Salary is a continuous variable, we used Regression algorithms.
- We used average salary of all categorical as Baseline Model
- The models used are : Linear Regression, Extra Trees Regression and Light GBM Regression
- For Extra Trees Regressor and Light GBM Regressor, we tuned hyper parameter using RandomizedSearchCV to reduce error.
- We then used 5 fold Cross Validation to find the best model

Model Evaluation

- We used Mean Squared Error as evaluation metric. It is the average squared difference between the estimated values and the actual value.

Model Performance Results

- Baseline - 644.2563
- Linear Regression - 358.148420
- Extra Tree Regression - 316.868358
- Light GBM regression - 307.094494
- The Light GBM regressor performed best on the problem with MSE of 307.0944 as compared to baseline model having MSE as 644.2563.
- Mean squared error was improved by approx. 52% on baseline model.

Analysis Results & Recommendations

Result #1

- As predicted by Light GBM Regressor, the ‘Miles from Metropolis’ is the key predictor, followed by Years of Experience.
- Other variables like group mean absolute deviation, group maximum, group minimum, group median, group mean, industry, job Type, major, company id and degree also had considerable role in predicting salary.

Result #2

Salary varies according to the following factors:

- Salary decreases linearly with miles away from city.
- Salary increases linearly with years of experience.
- Job position: CEO > CTO, CFO > VP > Manager > Senior > Junior > Janitor
- Oil and finance industries are the highest paying sectors, while service and education are the lowest paying

Result #3

- Based on the analyses we can recommend that years of experience and location highly influence the salary, and Oil and Finance industries have highest salary even for entry level positions.
- Based on all of this information, companies can estimate the salary of new hire considering all factors, also candidates can decide the type of industry, location etc. to achieve the desired salary.

Next Steps & Improvements

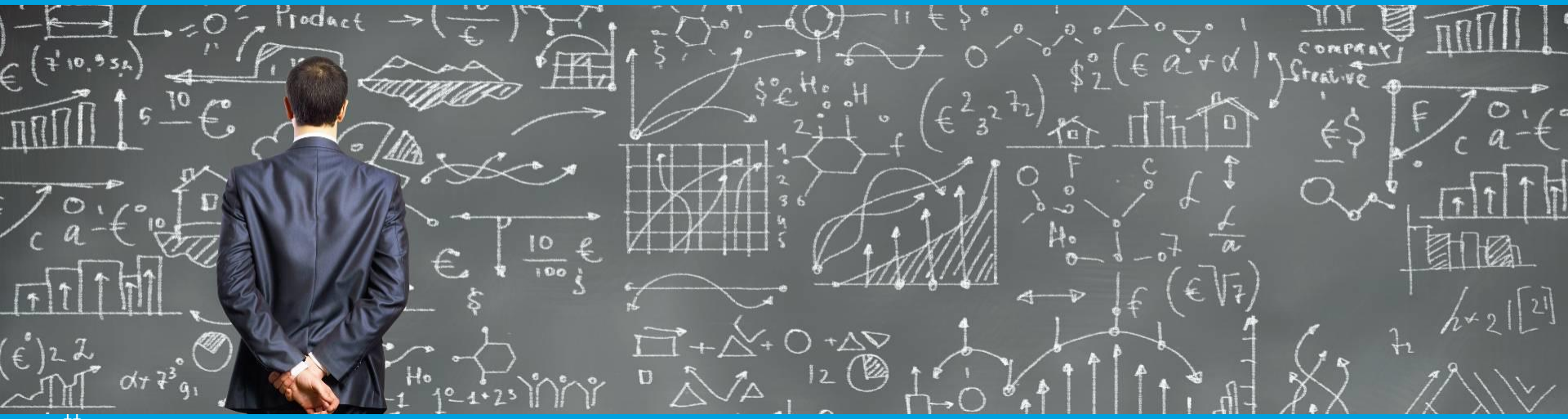
Project/Approach Improvements

1. We can further try to improve the model by engineering features from ‘years of experience’ and ‘Miles from Metropolis’.
2. We can test additional methods like Normalization and one hot encoding scheme.
3. If data had more properties like , ‘Type of employment’ (‘Full Time’ or ‘contractual’), ‘Overtime’, then model could have predicted salary more precisely.

Lessons learned

1. New engineered features can really boost the model’s performance.
2. Distance from the city is biggest factor in determining the salary. The more you live near city, more salary you can get.
3. Years of Experience is also a good factor

Appendix



Data Science Approach

1. Understand the problem	▪ Never forget which business problem you are trying to solve and the business objectives.
2. Explore the data	▪ Exploratory data analysis to understand the quality of the data (i.e. missing fields), the shape of the data (size, number of features, type of features), the statistic profile of the data (i.e. outliers, distribution etc.)
3. Cleanse the data	▪ Clean any data quality issues: garbage in, garbage out
4. Preprocess the data	▪ Transform the data or engineer new features if necessary to gain more insights
5. Metrics and Modeling	▪ Model creation, evaluation and selection
6. Evaluate findings	▪ Are they logical and do they make sense? Is the modeling approach used appropriate?
7. Iterate and Refine	▪ Refine analysis and fine tune models and findings
8. Communicate clearly	▪ Simple and straightforward messaging linking the results to the business outcome. ▪ Assumptions stated.

Development Environment

Code

https://github.com/samitsingh/Salary_Prediction

- Python
- Libraries: Scikit-Learn, Pandas etc.

Environment

- Windows, Linux
- Anaconda
- JupyterLab

Visualization

- Seaborn,
- Matplotlib

Data

- CSV