Salary Predictions

Contents

Business Problem

- The problem addressed in this project entails how HR department can offer reasonable salary to staff and thereby be able to reduce unnecessary cost to the company while maintaining positive employee motivation.

- If staff happens to be underpaid, there will be high employee dissatisfaction resulting in increased employee turnover within the company. While on the other hand, overpaying can increase company's cost which could have been used to further the growth and expansion of the company.

Key Takeaways

- ‘Job Type’ and ‘Degree’ are found to have a high impact on the target variable, ‘Salary’
- The respective roles were further grouped by ‘Job Type’, ‘Degree’, ‘Major’, ‘Industry’ resulting in few features
- Consequently, it is seen that the ‘mean’ of each group has the highest impact on the ‘Salary’

Dataset Information

- The Dataset comprises seven features which can help us determine the salary
- 'Salary' is the target variable we are predicting

Categorical Features

Job Id

Company Id

Job Type,

Degree,

Major

Industry

Numerical Features

Years of Experience

Miles for Metropolis

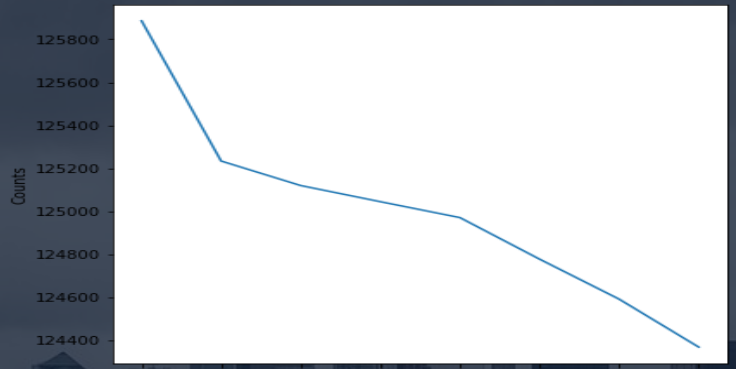
Salary

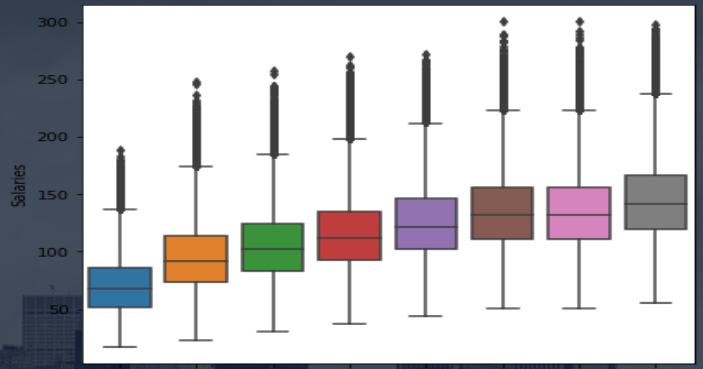
“The best way to predict the future is to create it”

Peter Drucker

Exploratory Data Analysis /EDA

Relationship between Target Variable (Salaries) and Input Variables

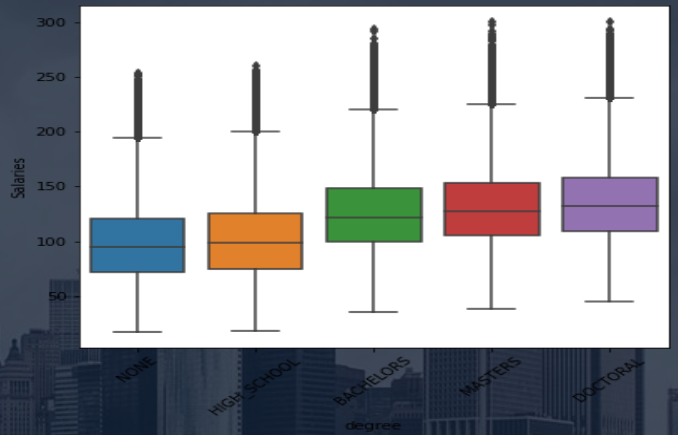




Salary is positively correlated with job type

Exploratory Data Analysis /EDA

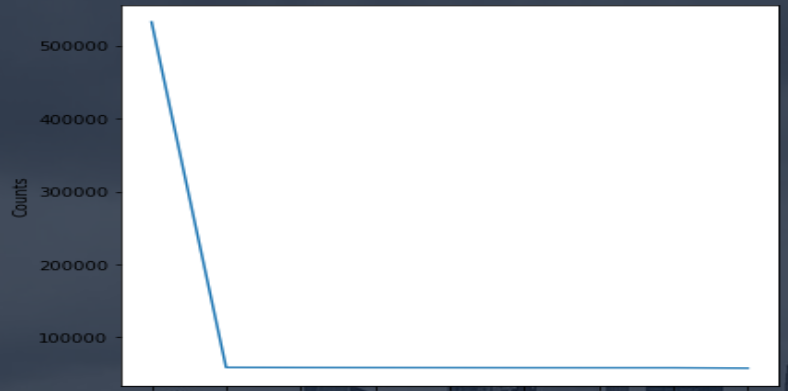
Relationship between Target Variable (Salaries) and Input Variables

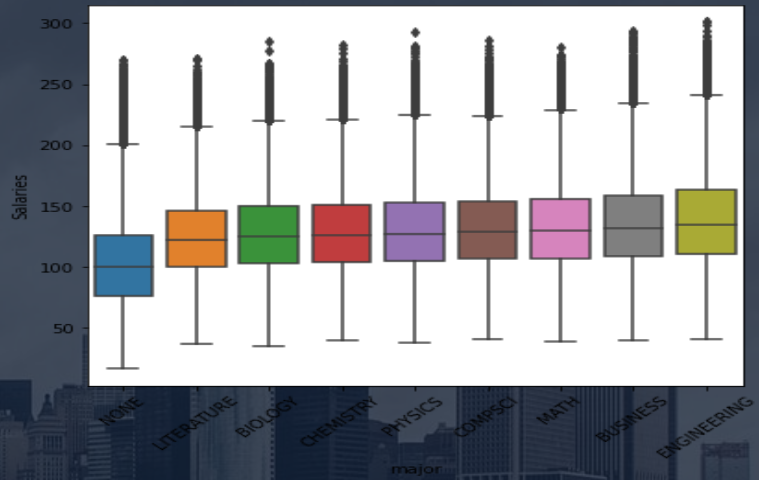


High salaries tend to correspond to advanced degrees

Exploratory Data Analysis /EDA

Relationship between Target Variable (Salaries) and Input Variables

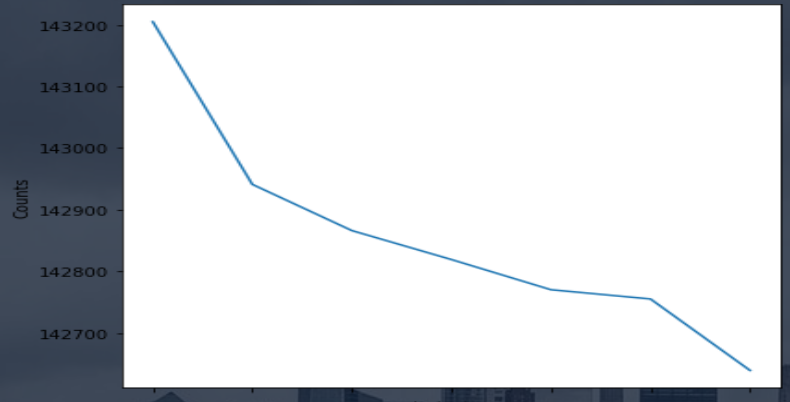


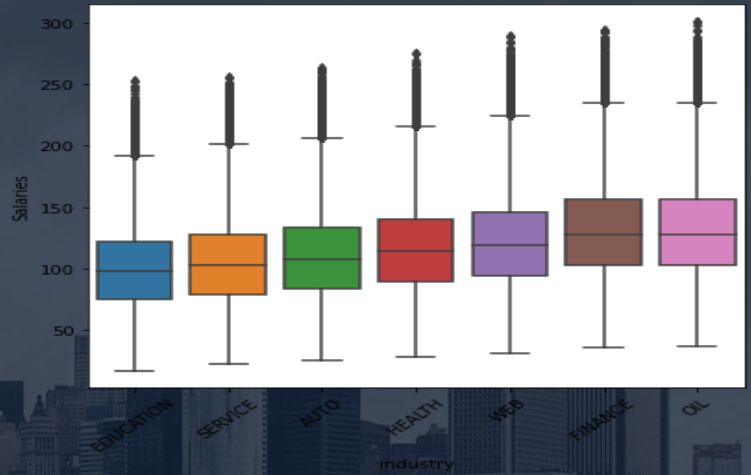


Employees with majors of engineering, business and math have corresponding high salaries

Exploratory Data Analysis /EDA

Relationship between Target Variable (Salaries) and Input Variables

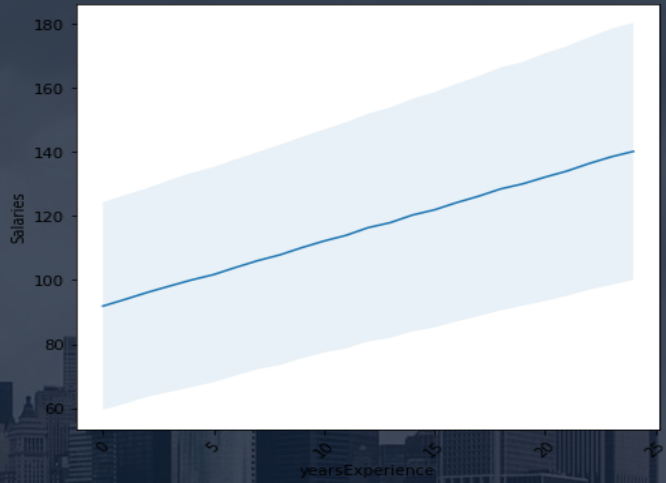




Higher salaries correspond to the oil, finance and web industry sectors

Exploratory Data Analysis /EDA

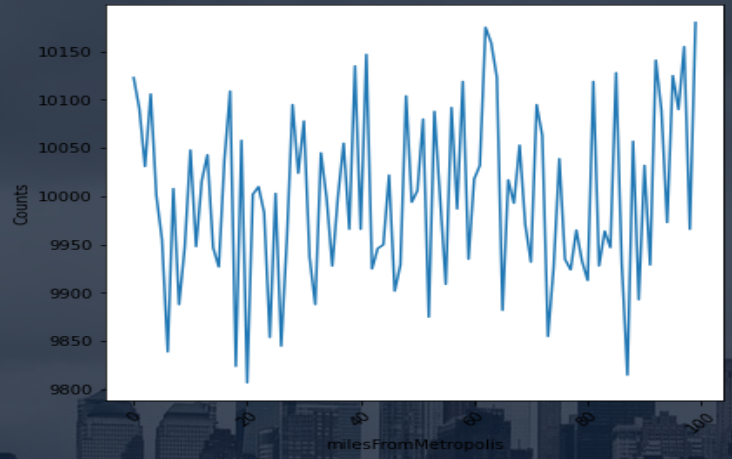
Relationship between Target Variable (Salaries) and Input Variables

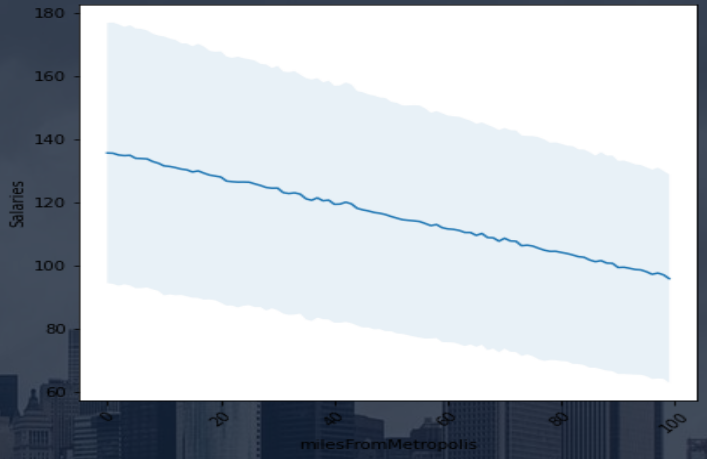


Salary is directly correlated with years of experience

Exploratory Data Analysis /EDA

Relationship between Target Variable (Salaries) and Input Variables

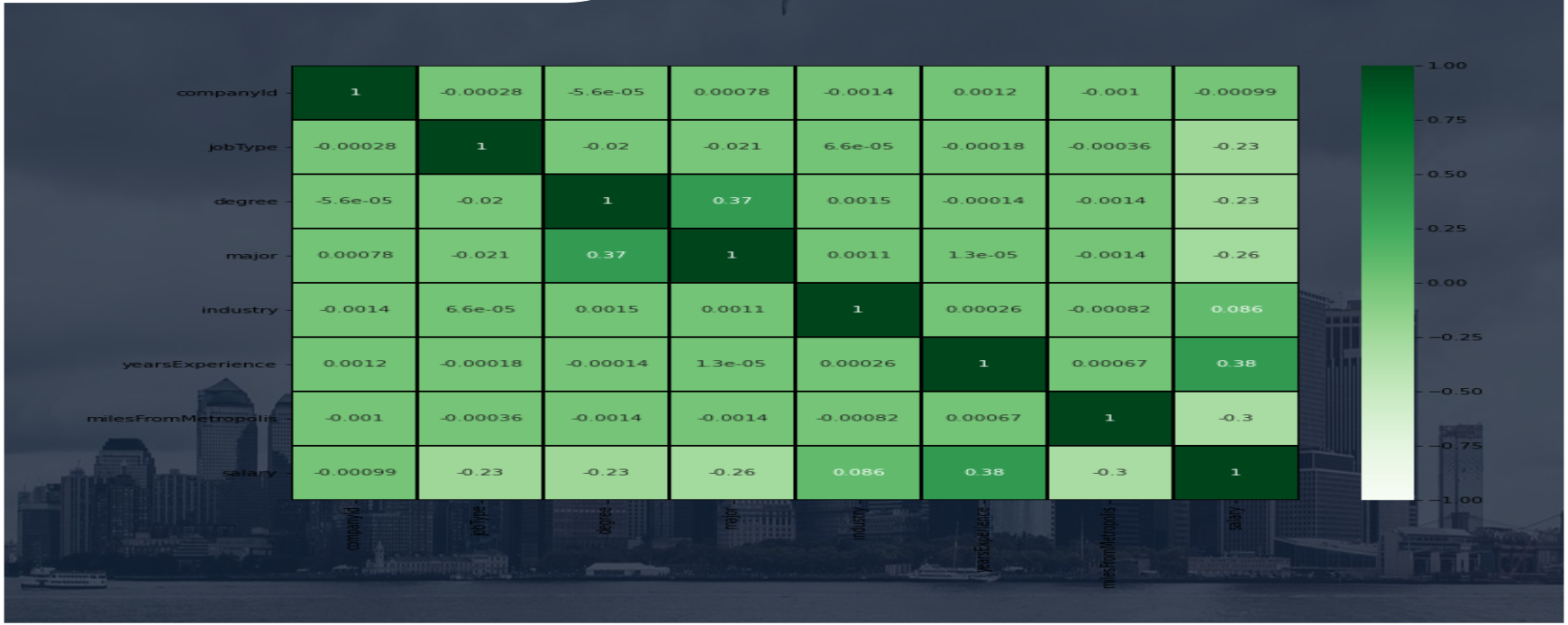




Salaries decrease as you go further away from the metropolis

Exploratory Data Analysis /EDA

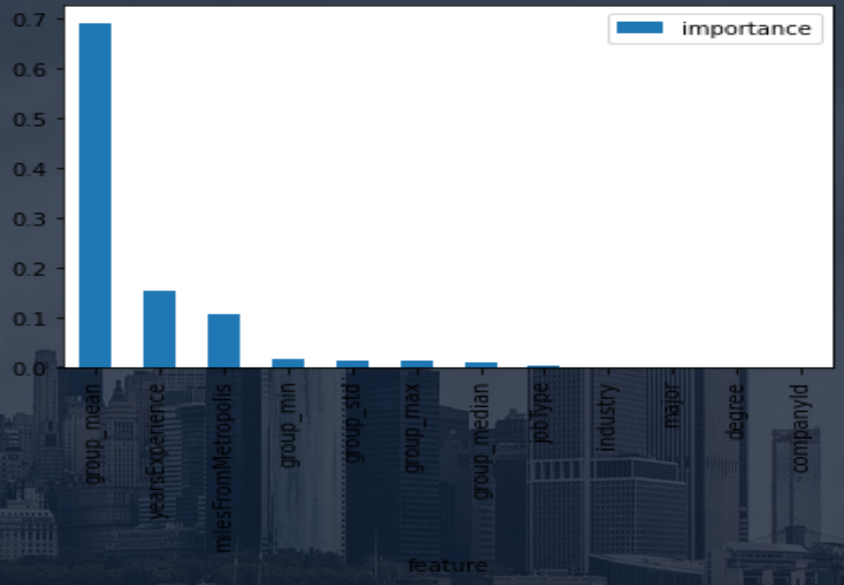
Visualizing the Correlation across all Columns



Data Preprocessing and Feature Engineering

Categorical features were grouped and new statistical features were calculated for the 'Salary' for each group

- ✓ group_mean
- ✓ group_std
- ✓ group_max
- ✓ group_min
- ✓ group_median



Modelling and Evaluation

- Supervised Machine Learning algorithms, namely, Regression and Ensembles of Regression Algorithms suit our data and efficacy goal.

- Three models were chosen:

  - ✓ Linear Regression : sometimes simple is best
  - ✓ Random Forest Regressor : offers Improved accuracy and control over-fittings
  - ✓ Gradient Boosting Regressor : can optimize on Least squares regression.

Model Results

The Efficacy Metrics is :

Mean Square Error / MSE

- Baseline Model : MSE 1499.00
- Linear Regression : MSE 358.16
- Random Forest Regressor : MSE 313.77
- Gradient Boosting Regressor : MSE 313.14

- Hence forth Gradient Boosting Regressor is chosen as the BEST MODEL

Next Steps

- There is window of opportunity to further explore and improve the model by creating new features in relation to 'miles from metropolis' and 'years of experience' features.
- File is saved appropriately for further testing / deployment.


Thank You