Salary Predictions

Ovo Okpubuluku
Calgary, AB, Canada

June
2020

Contents

Overview

Salary Predictions based on Job Descriptions

- Our goal in this project is to examine this dataset of job postings, and predict salaries for a new set of postings.
- This will involve building a model to predict the salaries given in the test dataset.

Defining the Goal

Practical uses of this prediction

- HR Department of a large company or a Consulting Outfit that needs real-time solutions in order to make effective employment offers to potential hires.
- To understand current realities in the job market and how businesses can leverage this in order to secure high quality talent, whilst keeping hiring costs low.

Dataset Information

- The Dataset provided contains 7 features which can help us determine the salary
- Salary is thus identified as the target for our prediction

Categorical Features

- jobId - Unique for each entry - categorical
- companyId – Categorical - 63 companies
- jobType – Categorical – 8 categories
- Degree – Categorical – 5
- Major – Categorical – 9 categories
- Industry – Categorical – 7 categories

Numerical Features

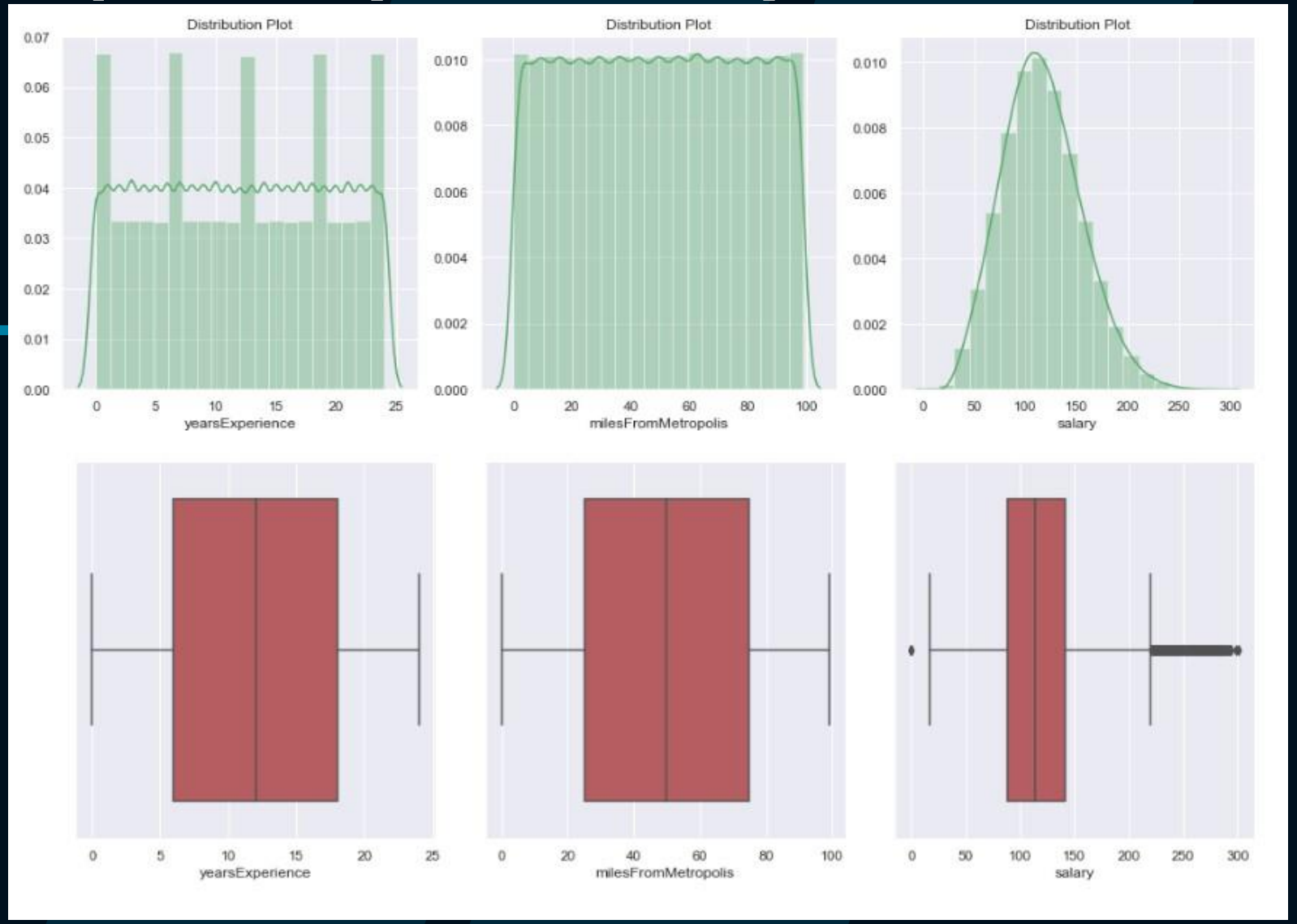
- yearsExperience
- milesFromMetropolis
- Salary(Target)

Exploratory Data Analysis

Overview of Train Dataframe showing Features and Target

	jobId	companyId	jobType	degree	major	industry	yearsExperience	milesFromMetropolis	salary
0	JOB1362684407687	COMP37	CFO	MASTERS	MATH	HEALTH	10	83	130
1	JOB1362684407688	COMP19	CEO	HIGH_SCHOOL	NONE	WEB	3	73	101
2	JOB1362684407689	COMP52	VICE_PRESIDENT	DOCTORAL	PHYSICS	HEALTH	10	38	137
3	JOB1362684407690	COMP38	MANAGER	DOCTORAL	CHEMISTRY	AUTO	8	17	142
4	JOB1362684407691	COMP7	VICE_PRESIDENT	BACHELORS	PHYSICS	FINANCE	8	16	163
5	JOB1362684407692	COMP15	MANAGER	DOCTORAL	COMPSCI	FINANCE	2	31	113
6	JOB1362684407693	COMP15	CFO	NONE	NONE	HEALTH	23	24	178
7	JOB1362684407694	COMP24	JUNIOR	BACHELORS	CHEMISTRY	EDUCATION	9	70	73
8	JOB1362684407695	COMP20	JANITOR	HIGH_SCHOOL	NONE	EDUCATION	1	54	31
9	JOB1362684407696	COMP41	VICE_PRESIDENT	BACHELORS	CHEMISTRY	AUTO	17	68	104

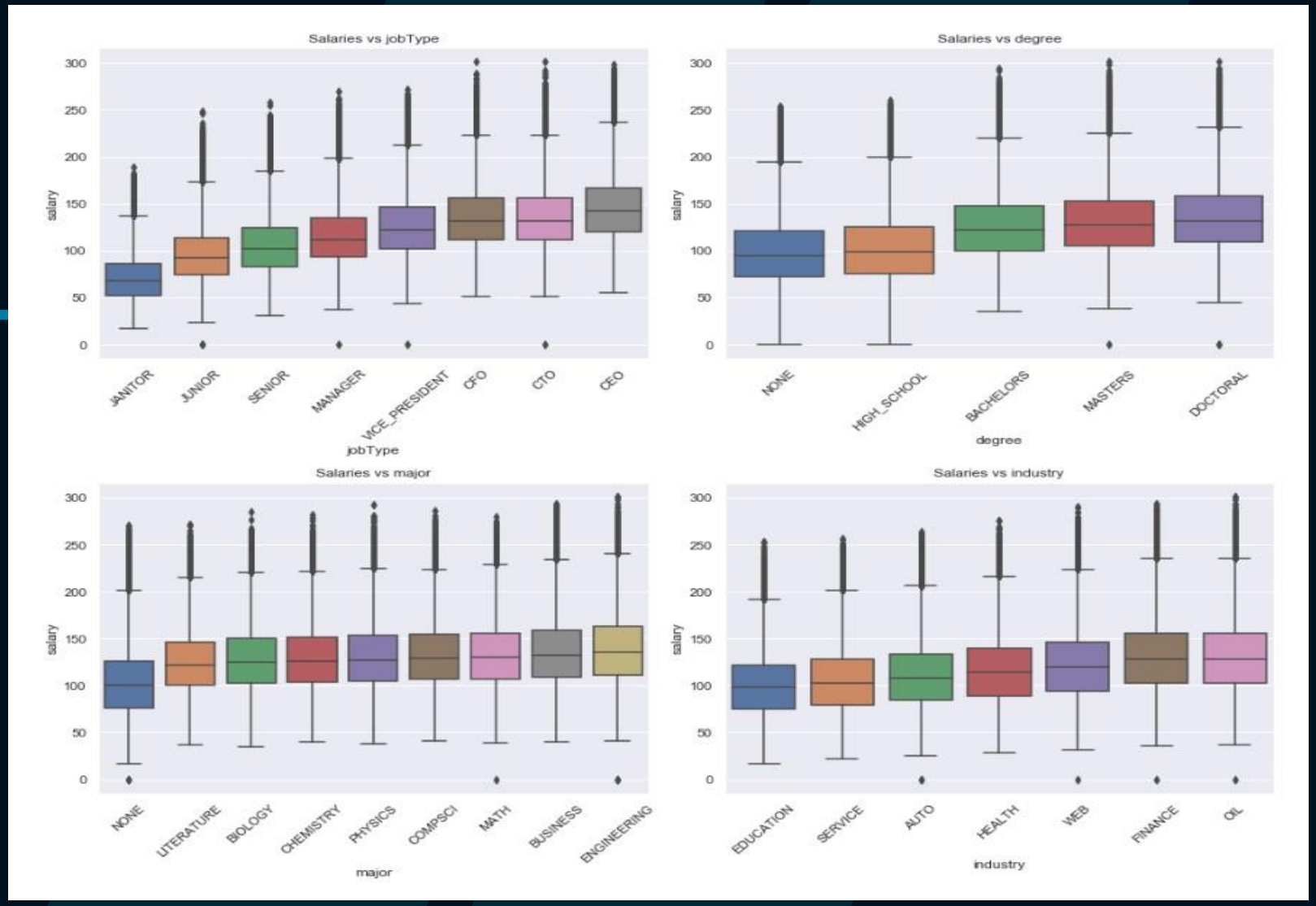
Exploratory Data Analysis



Visualizations for numerical features

- We can see the presence of outliers in the Salary boxplot.
- After close examination, the zero values look unrealistic for those positions. There were 5 of them.
- The outliers in the upper range are realistic for senior level salaries

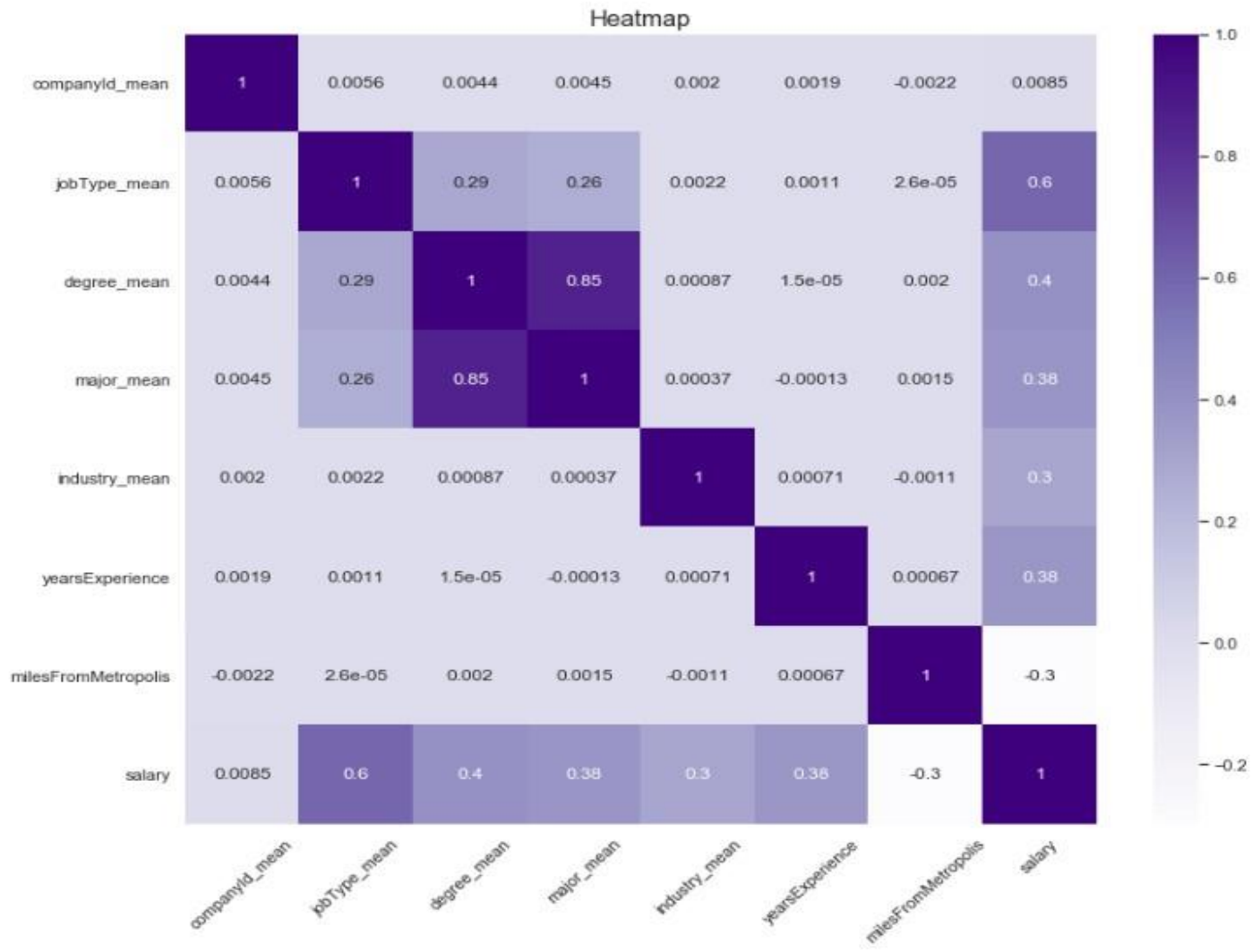
Exploratory Data Analysis



Visualizations for categorical features

- The Senior Job types obviously earn the highest salaries (Strong positive correlation)
- Advanced Degrees tend to attract higher salaries
- Engineering, Business, Math and Computer Science are on the top end of the salary continuum
- The Oil and Finance Industries are the highest paying

Exploratory Data Analysis



Correlation Matrix (Heatmap)

- Categorical features were converted to numeric by label encoding
- jobtype as the most strongly correlated feature with salary.
- Degree and Major have the strongest positive relationship

Feature Engineering and Data preprocessing

Categorical features were grouped and group statistics selected as new features:

- mean salary
- Median salary
- Minimum salary
- Maximum salary
- Mean absolute deviation salary

MODEL SELECTION & EVALUATION

We selected three different Regression Algorithms for Evaluation:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor

Mean Squared Error was used as evaluation metric

MODEL RESULTS

Baseline Model: 644.26

Linear Regression: 358.15

Random Forest Regressor: 313.27

Gradient Boosting Regressor: 313.06 (selected as best model)

Key predictors for this model is the group mean salary, followed by yearsExperience

NEXT STEPS

- Files Saved for further testing/deployment
- We could seek to further improve the model by creating new features around the yearsExperience, and miles from metropolis features

THANK YOU!

Ovo Okpubuluku

Phone
403-397-0259

Email
ovo@live.ca