MIS S380N - Group Project

By: Apurva Audi, Amanda Nguyen, Shubhada Kapre, Victor Lai

Description of the Problem

Description

This dataset chosen includes 62,000 salary records from top companies. We were looking for patterns in certain features that could help predict salaries for data science and STEM professions. Some predictor variables included job location, years of experience, gender, and level of academic experience (high school, undergraduate degree, doctorate, etc.). Additionally in this data set, we selected a number of specific questions to answer. These questions included:

1. What are the highest paying jobs in the tech industry?
2. Which companies pay the highest salaries?
3. Do men earn more than women in the tech industry?
4. Where are the tech jobs located?

Importance of problem

This dataset was special as it applies to all students in the current MSBA program both at The University of Texas at Austin and at other colleges. The dataset provides useful information about top companies across a wide span of industries from June 2017 to August 2021. We picked this problem because we wanted to find out if there were certain predictors that could provide indication on increased salary level. Beyond the scope of masters students, it applies to anyone who is trying to enter the technology market for work. However, our findings can also be applied to different industry and occupation types and replicated for further salary investigation.

Students wanting to enter into FAANG companies or large data science and stem professions would appreciate the analysis of these problems. From the information gathered in our findings, one can direct their career choices based on specific companies, locations, and schooling experience.

Exploratory Analysis

Univariate Analysis

Overall, each of the variables are not very well balanced. Many numerical variables had an exponential distribution, which could potentially make modeling difficult. Several of the features were also redundant and had missing values.

Data Pre-Processing

To correct the exponential distributions, the features regarding total yearly compensation, years of experience, and years at company were log transformed. This reduced the number of points initially perceived to be outliers and made the overall distribution more gaussian in appearance. Redundant columns were simply dropped from the data, and missing values were replaced with an "Other" category

Patterns or Abnormalities Found

Years of experience and years at company as well as presence of a doctorate degree have a positive correlation with compensation, which is more or less expected. Strangely, having a bachelor's degree is negatively correlated with compensation, while having a master's or doctorate is weakly positively related. Also abnormal is the fact that every race seems to have negative correlations, with not one race seeing benefits. [Figure 1]

What are the highest paying jobs in the tech industry?

The highest (5) paying jobs in the tech industry include Software Engineer Manager, Product Manager, Technical Program Manager, Hardware Engineer, and Solution Architect. [Figure 2]

Do men earn more than women in the tech industry?

Yes, by about 3.3% on average. [Figure 3]

Which companies pay the highest salaries?

Based on our analysis of companies with more than 500 employees, the highest annual median salaries in the tech industry range from $130k - $290k. Facebook offers the highest annual median salary followed by Uber, LinkedIn, Apple,Google and Salesforce in the same order. [Figure 4]

Where are the tech jobs located?

The bar plot shows the top 5 locations for tech jobs. Seattle, WA has the highest number of tech jobs with over 8000 employees in the tech field. [Figure 5]

Solutions and Insights

Ordinary Least Squares

When running the ordinary least squares regression, each feature is deemed important given p-values less than 0.05, indicating statistical significance at the 95% confidence level. However, the R-squared of the model is quite low (0.3659), meaning that it may not be the most accurate representation of the true population. Due to numerous categorical variables, the creation of dummy variables originally produced over 4000 rows. In order to minimize this, the top 5 largest locations, companies, and job titles were kept, while the remaining predictors in each category were clustered together into an ’other’ column. From there, all features were used to ensure significance from the linear regression. As a result, it was determined that it may be better to run additional models to find one that fit the data better.

Ridge Regression

When running ridge regression, the first step included finding the correct tuning value for lambda. All values between .001 and 1 were tested to produce the ultimate shrinking factor hyperparameter of 0.001526. Then using that value and applying it in the ridge regression to both training and test data, the output values were as followed:

- RMSE: 0.4696 (training set) & 0.4728 (test set)
- R-Squared: 0.4118 (training set) & 0.4179 (test set)

Feature importance: we were able to get a general idea of which features were important, years of experience seeming most influential.

Lasso Regression

Similar to ridge regression, the first step for running lasso regression involved discovering the proper tuning parameter. Upon testing values between 0.001 and 1, the grid search returned a best lambda value of 0.001. After applying the value to the Lasso model, the following returned:

- RMSE: 0.5313 (training set) & 0.5371 (test set)
- R-Squared: 0.2471 (training set) & 0.2493 (test set)

Feature importance: using a lambda of .001, all of the coefficients shrunk to a value of zero.

Nearest Neighbors

K Nearest Neighbors calculates the distance to the nearest K neighbors to predict the target variable and hence takes a long time to run on large data sets. To make the computation faster, a random sample containing 30% observations of the original data was taken to run KNN with all the features as input variables and the total yearly compensation in dollars as the output target variable. The RMSE was calculated by performing a 3-fold cross validation on the data with K value ranging from 1 to 300. The obtained RMSE was plotted against the K value. The plot showed that the least RMSE of 321.78 is obtained by using a K value of 30. [Figure 6]

RandomForest

Random Forest is extremely slow when the number of features increases to thousands as a result of one hot encoding, so first the timestamp column was converted to year, location was converted to state/country, and other categories were converted to “Other” if they had less than 500 data points and so applied to less than 1% of the 62.6k total data points.

The main tuned parameters were the number of estimators, the maximum number of features tuned per split, and the minimum number of samples allowed per leaf. On the test set, the tuned model achieved a mean squared error of 0.097692, translating to an RMSE of 0.312, or around 36% error per compensation prediction due to the log scale.

The two most important features by far were years of experience and city id. The number of years spent at the company also held some amount of weight. Since random forest is hard to interpret, a simple linear regression was fit using only the top 10 most important features from random forest to try to identify the direction of each importance. [Figure 7]

Boosting

Gradient boosting builds an additive model by using multiple decision trees of fixed size as weak learners or weak predictive models and hence, ensuring creation of a more robust model with a higher predictive power.

All predictors of string data-types were converted into upper case and label encoding is used to generate labels for all unique string values. Additionally, ‘year’ from timestamp predictor is extracted as it is a more meaningful parameter to gauge correlation with yearly compensation. With a train-test dataset split of 70-30%, the initial Histogram Gradient Boosting Regressor with hyper parameters : number of boosting stages as 500, and learning rate of 0.01, achieved a R^2 score of 0.713 and RMSE of 0.3277.

However, it is essential to use only the most important variables for modeling and hence, relative feature importance was computed to account for relevant variables only. [Figure 8]

Results of Feature Importance:

- The most relevant variables are cityid, years of experience, company, location, level, title, tag, years at company, Doctorate and Bachelor’s degree in the same order.
- It is refreshing to know that gender and race don’t play any significant role in determining the salary of the individual in the tech industry.
- However, it is certainly surprising that a Master’s degree isn’t an important predictor for salary as compared to a Bachelor's or Doctorate degree.

Optimized Model : Using these selected features and RandomizedSearchCV for hyperparameter tuning, the best model is obtained with maximum boosting stages of 1000,maximum depth of the trees as 7 and learning rate of 0.1. This model obtained a R^2 score of 0.804 and RMSE of 0.2707.

Figures

Figure1

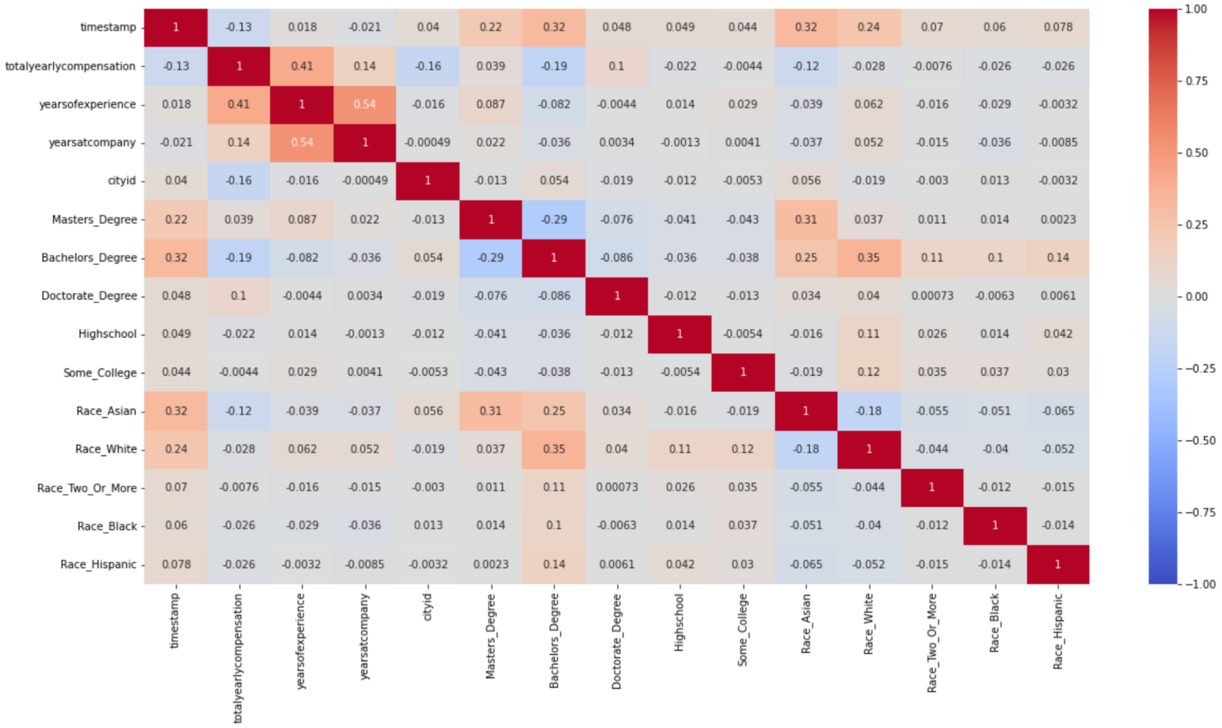


Figure 2

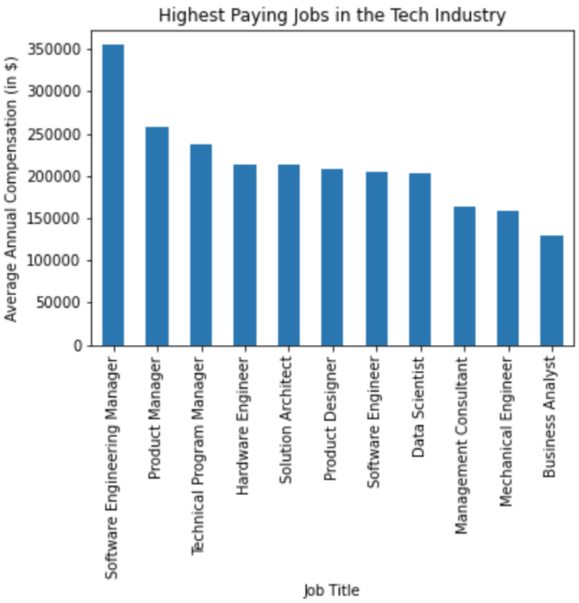
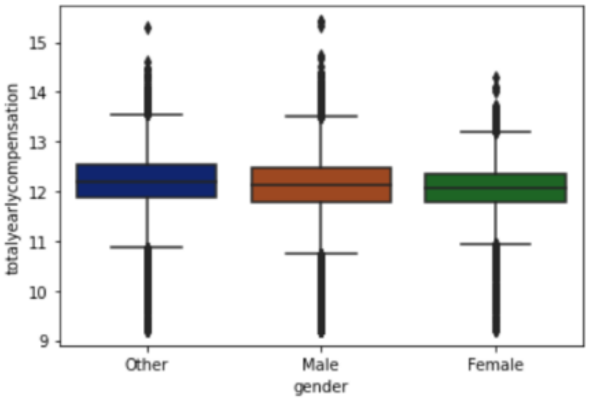
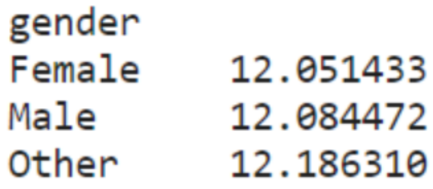


Figure 3





```
gender
Female     12.051433
Male       12.084472
Other      12.186310
```

Figure 4

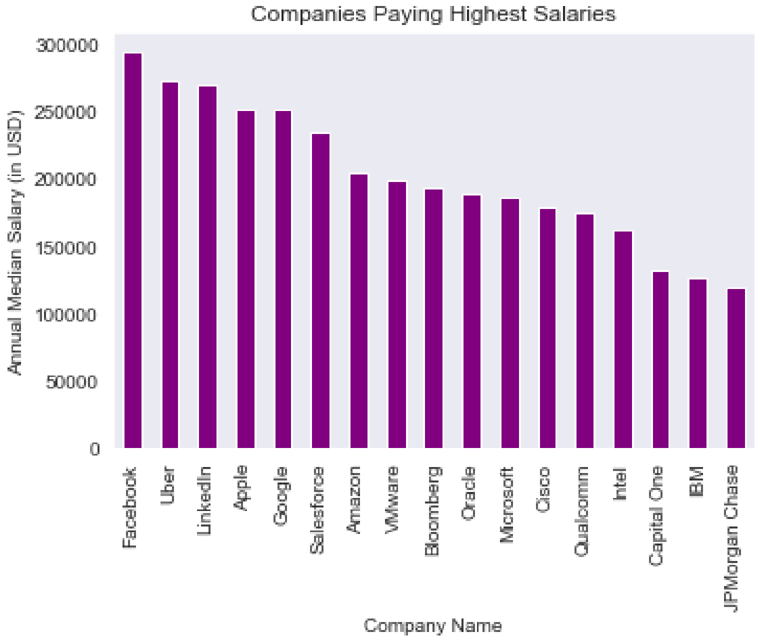


Figure 5

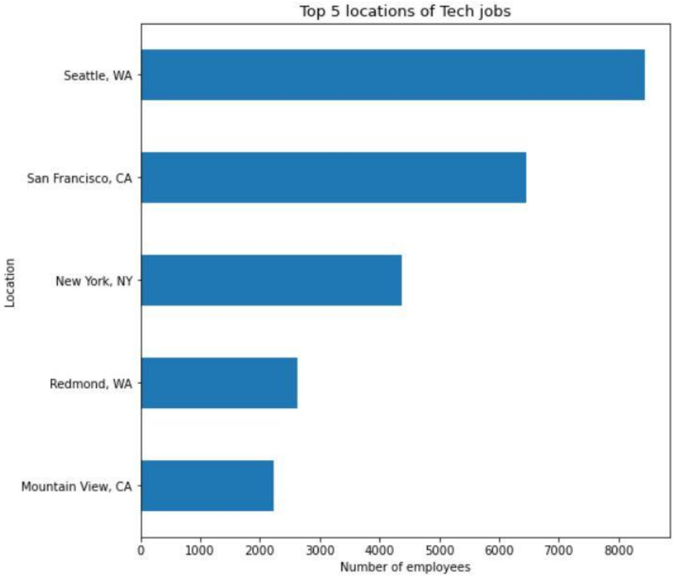


Figure 6

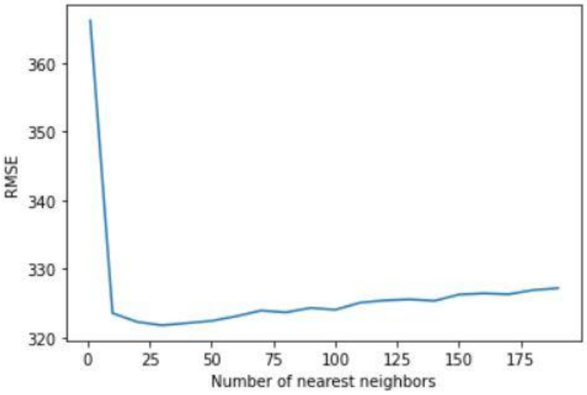
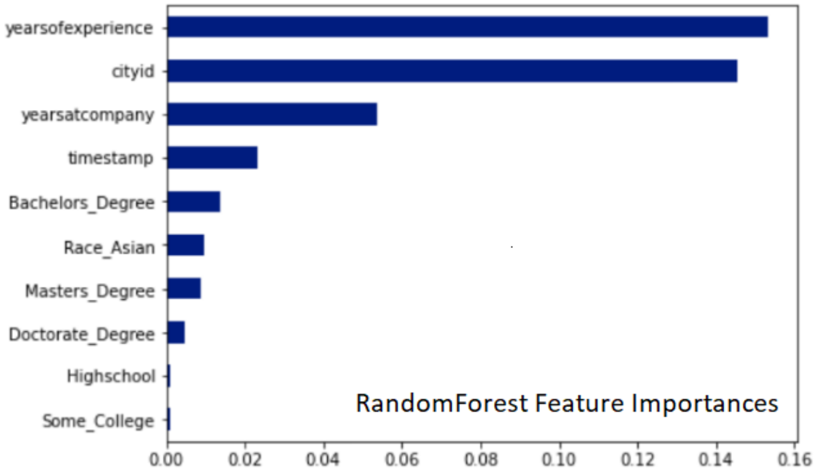
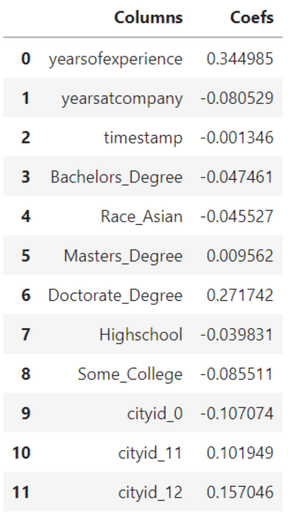


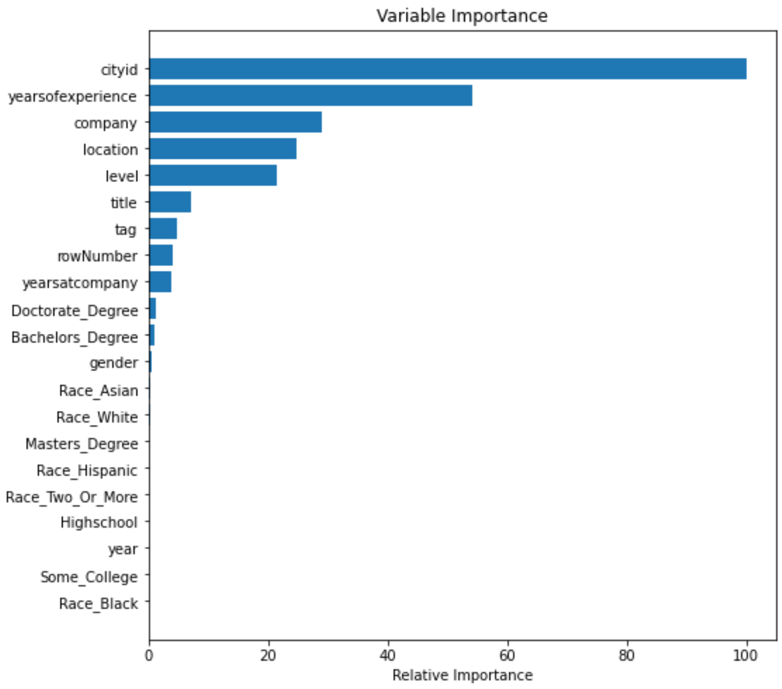
Figure 7





	Columns	Coefs
0	yearsofexperience	0.344985
1	yearsatcompany	-0.080529
2	timestamp	-0.001346
3	Bachelors_Degree	-0.047461
4	Race_Asian	-0.045527
5	Masters_Degree	0.009562
6	Doctorate_Degree	0.271742
7	Highschool	-0.039831
8	Some_College	-0.085511
9	cityid_0	-0.107074
10	cityid_11	0.101949
11	cityid_12	0.157046

Figure 8


Variable Importance
cityid
yearsofexperience
company
location
level
title
tag
rowNumber
yearsatcompany
Doctorate_Degree
Bachelors_Degree
gender
Race_Asian
Race_White
Masters_Degree
Race_Hispanic
Race_Two_Or_More
Highschool
year
Some_College
Race_Black
Relative Importance