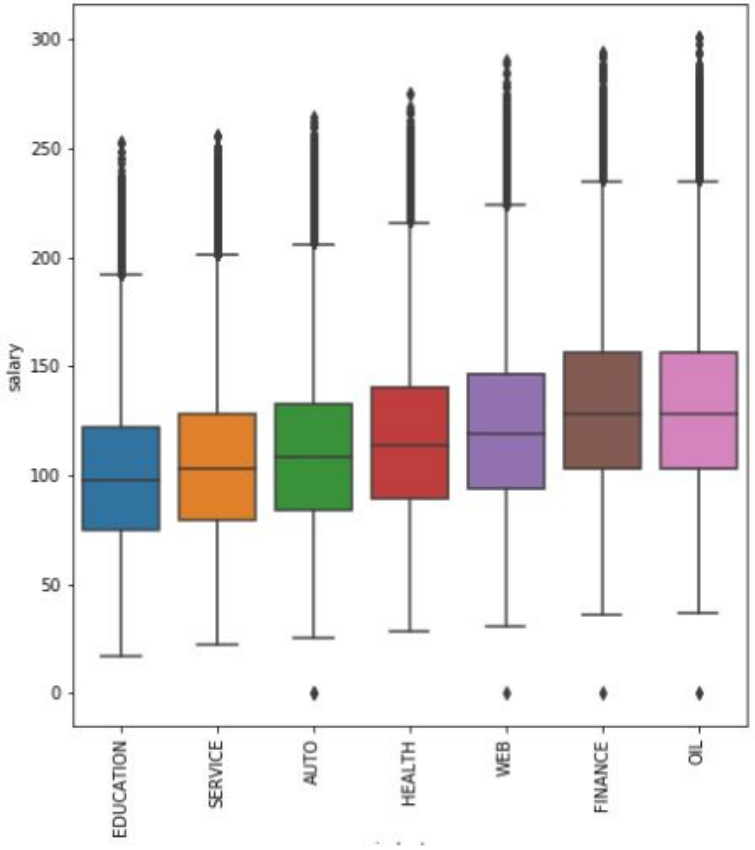
Salary Prediction Problem

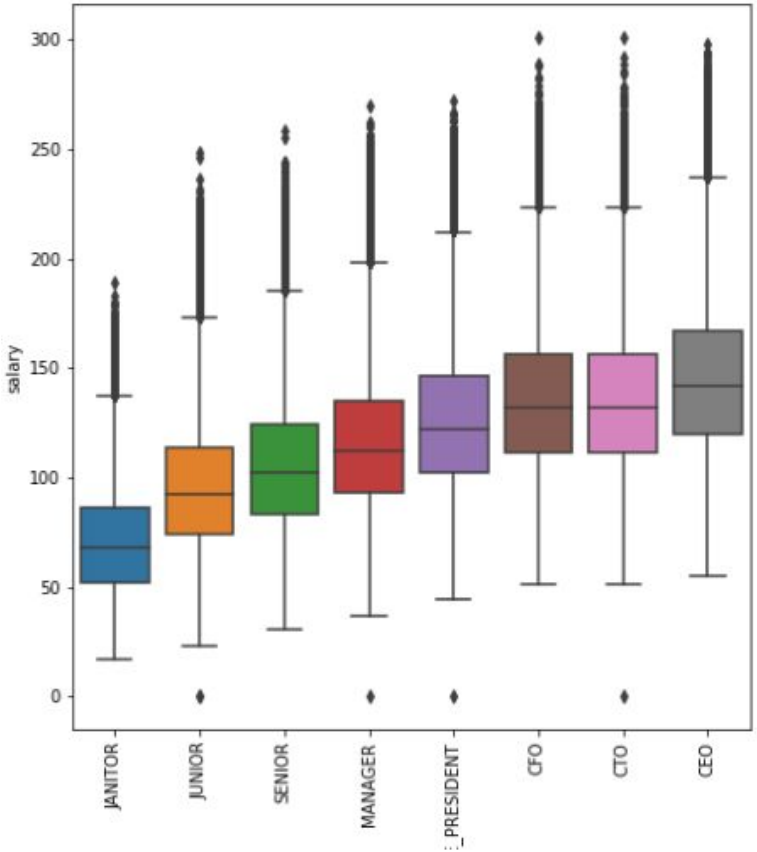
Define the problem

We would like to be able to predict salaries given certain features about a person: their experience, job type, education, and other factors.

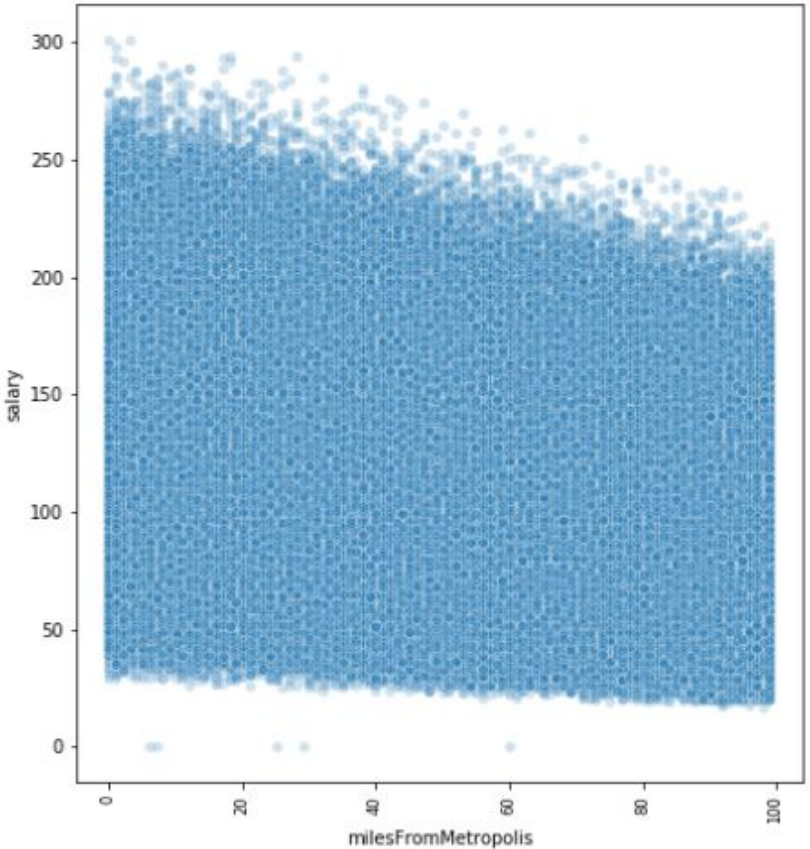
This is potentially valuable as it allows us to make sure we are not under or overpaying our employees. It could also help us define salary ranges for positions we are hiring for. Potentially highly valuable to recruiting agencies as they need to decide whether certain candidates are worth pursuing.

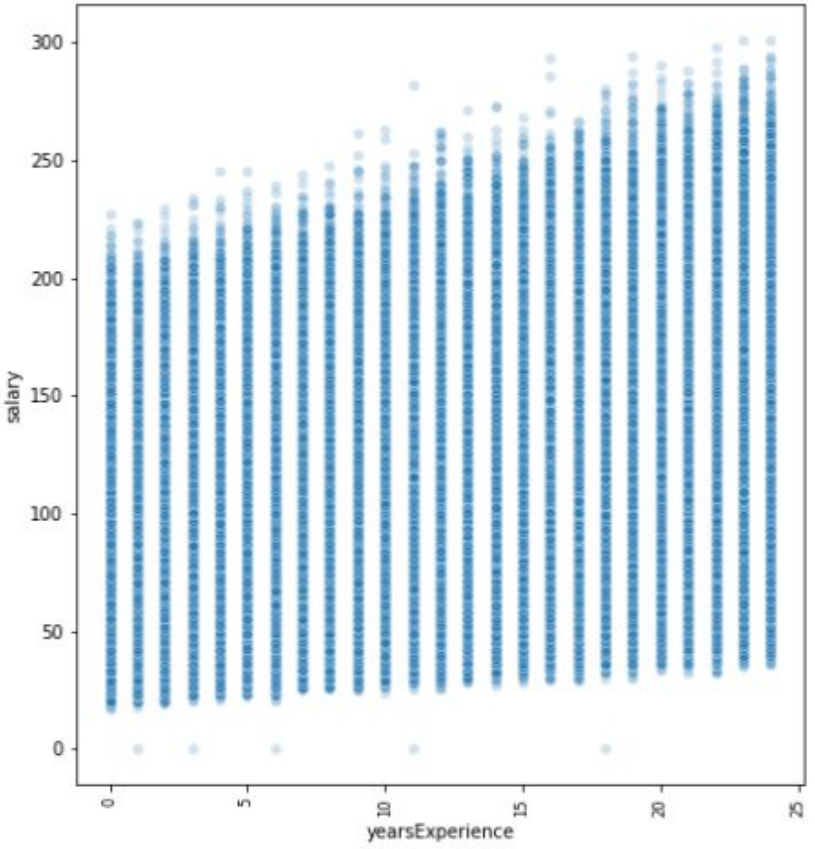
Explore the Data - Feature Exploration





Explore the Data - Feature Exploration



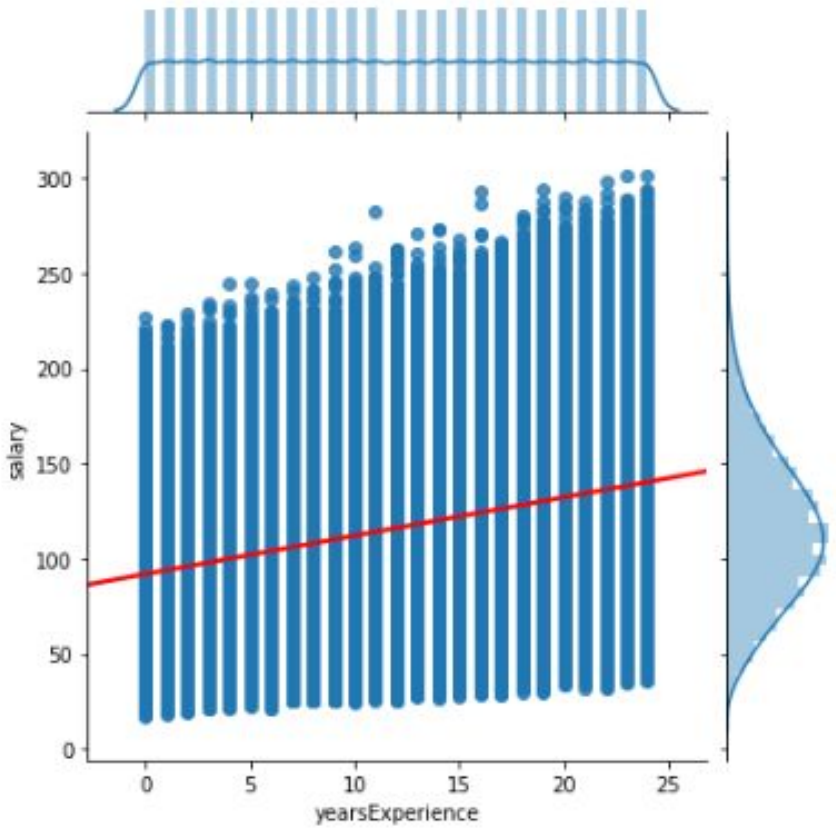


Explore the data - Correlation Matrix

- Most features are Correlated with salary According to the Correlation matrix and The previous plots
- This will be extremely Helpful as we try to Predict salary, and shows That the data is relevant To the task at hand.

Baseline

- Our baseline model is a linear regression Using the single feature of years of Experience.
- We will be using RMSE as our metric As it is easily interpretable and is a Natural choice for a quantitative Response variable
- The baseline model has an RMSE of 35.89. Since this is such a simple model, Our target should be to half this value.

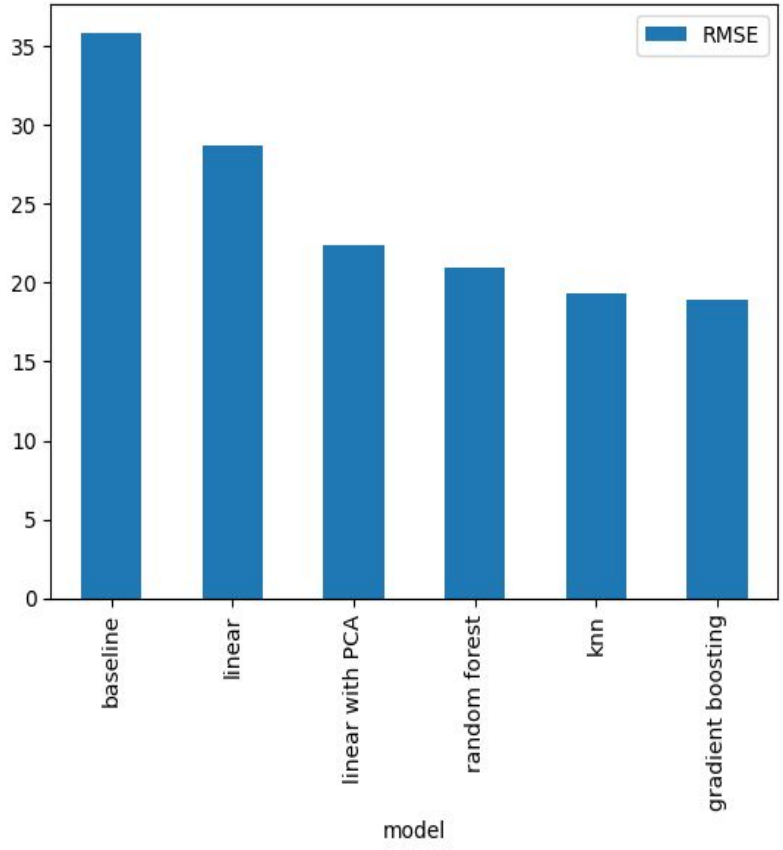


Potential Models

- We should attempt a linear model since many of our features appeared to have a linear relationship with the target.
- We should attempt a random forest and gradient boosted model as these typically have high performance.
- We will attempt a linear model after PCA to see if eliminating collinearity will help the linear model.
- Lastly, we will try a KNN algorithm to see if the spatial relationships between the features is exploitable.

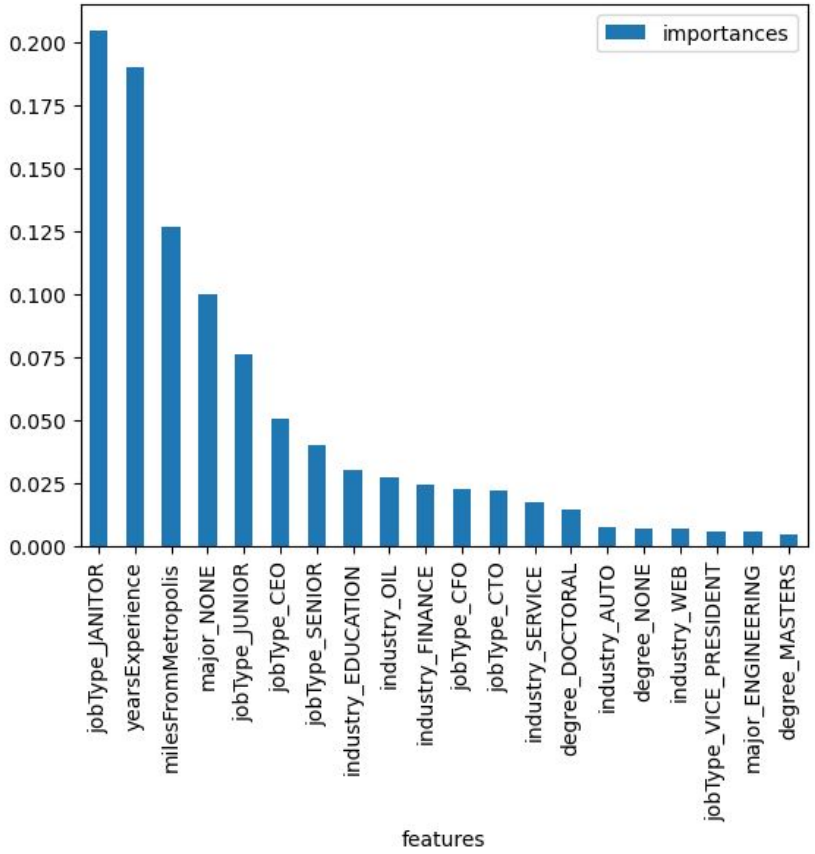
Model Performance

- Gradient boosting models performed best, With K Nearest Neighbors close behind.
- The gradient boosting model had an RMSE of 18.9, which is a 47.5% decrease from the Baseline RMSE of 36.
- The model significantly outperforms simple Linear regressions both with and without Feature scaling.



Feature Importances

- The most predictive feature appears to be Whether or not the job opening is for a Janitor position.
- Other highly predictive features are years Of experience, the number of miles the Employee lives from a metropolis, and Other specific job types.



Takeaways

- Baseline RMSE was 35.96, final model RMSE was 18.9. This shows an improvement of 47.5 %
- RMSE represents the standard deviation of our errors. This means that 68% of our predictions will be within 18.9k of the target, 95% of our predictions will be within 37.8k of the target, etc.
- Important features for predicting salary were years of experience, miles from metropolis, and the type of job in question.
- Knn could potentially be substituted for the gradient boosting model. Knn will have a much lower training time (none!) but will take longer to make predictions. This depends on the production needs of the model.