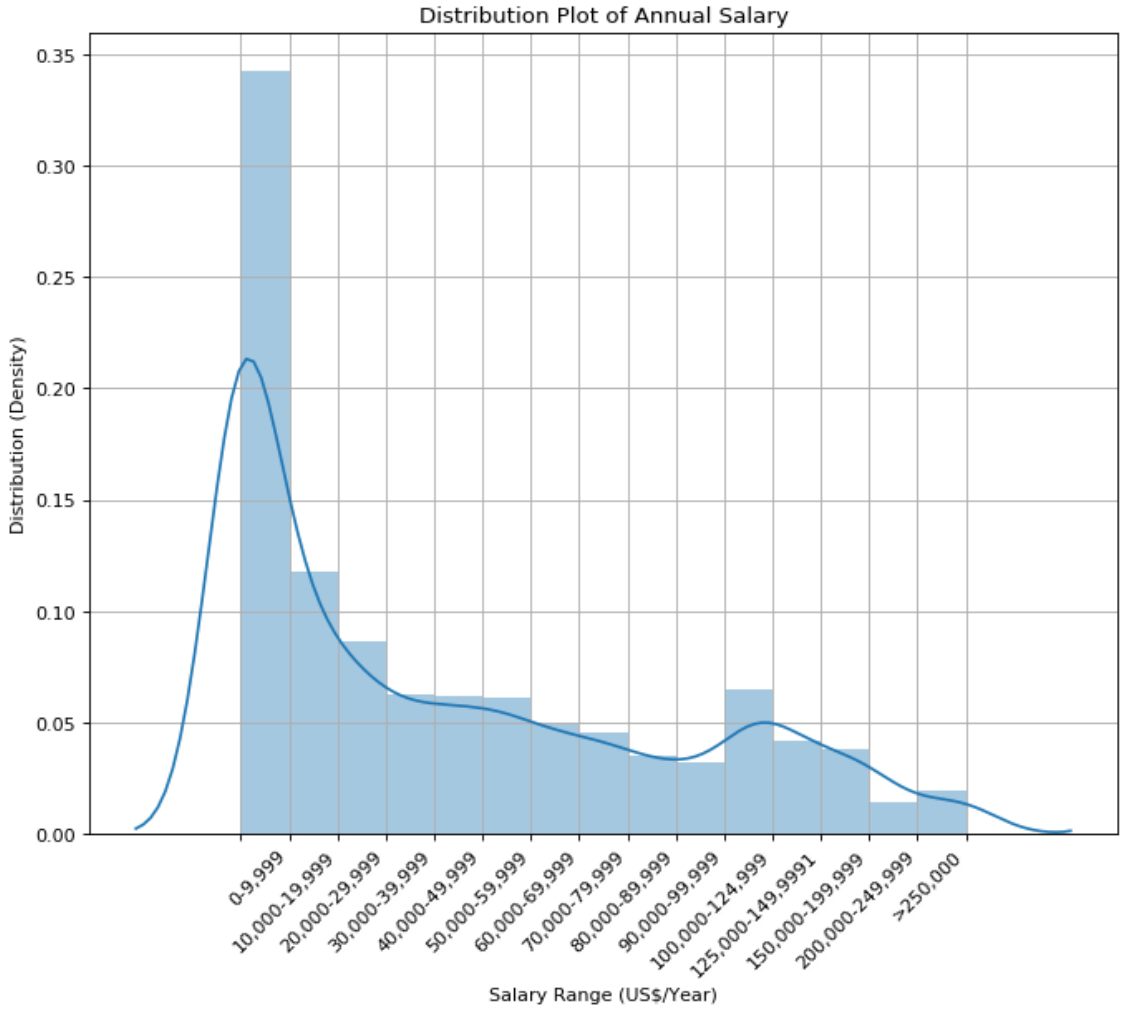
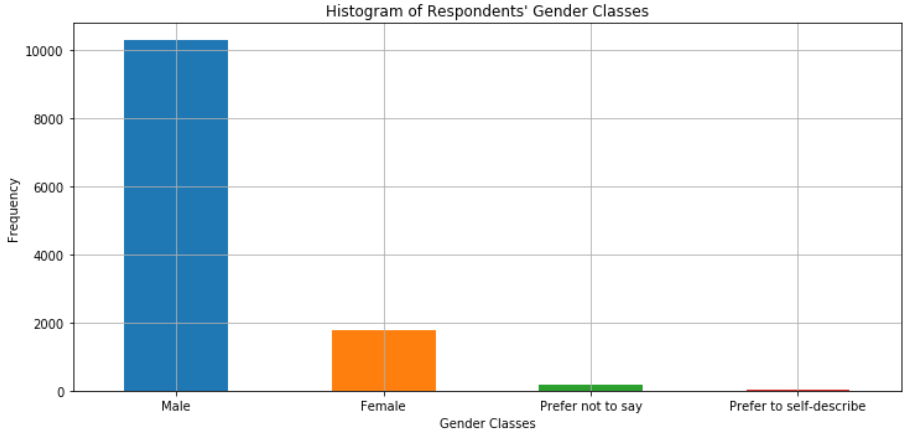
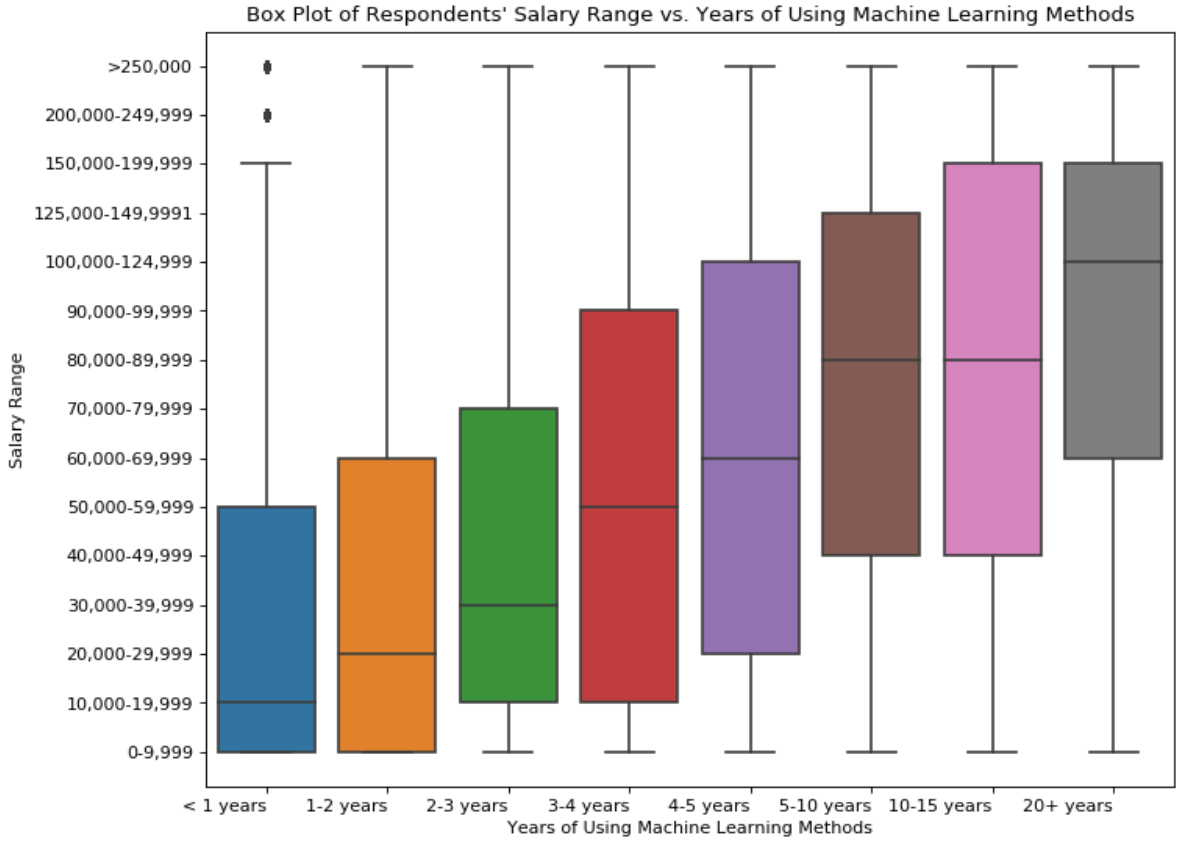
Findings from Exploratory analysis – Data set Structure

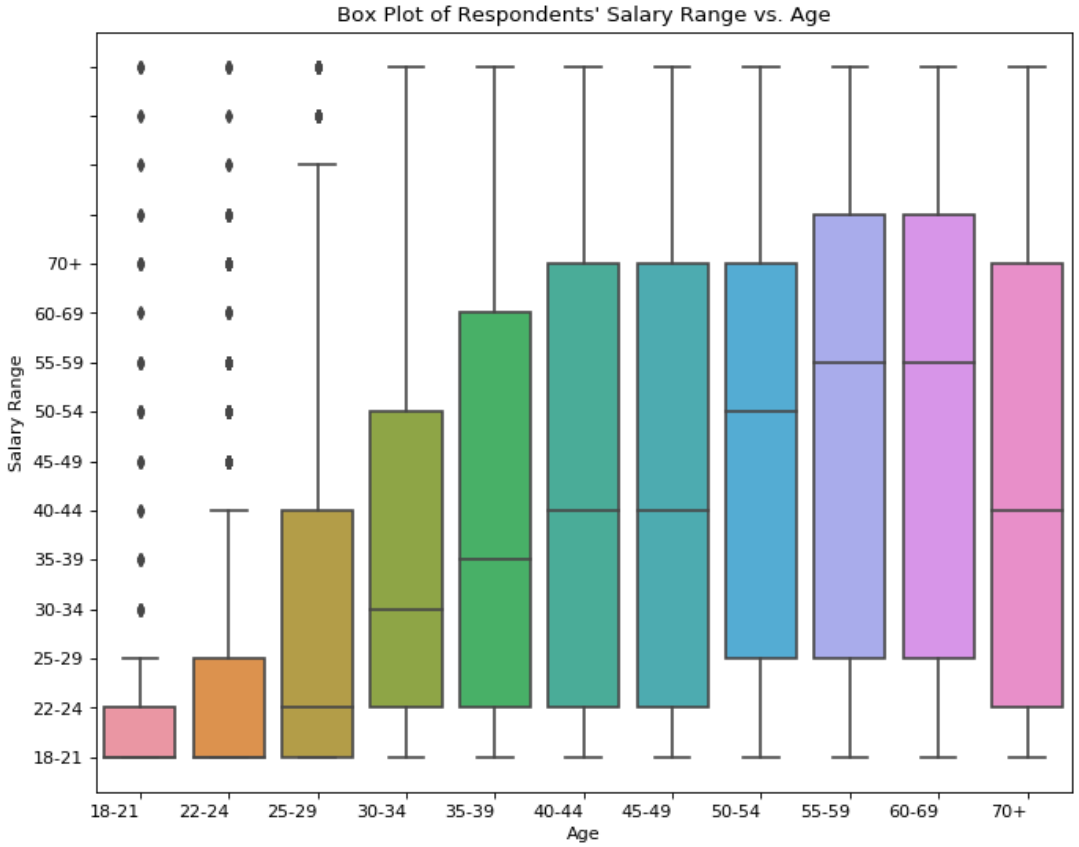




- As is shown in the salary distribution plot, the distribution is skewed heavily to the left with about 35% samples having an annual salary of less than $10000 per year. About 50% samples fail into the lowest three salary buckets (0-9999,10000-19999,20000-29999).
- This plot can be a useful reference to check if the later predicted targets also f all into a similar distribution
- As shown in the age and gender distribution plot, the majority of the respondents are aged the between 25-34 and there is a significant gap between the number of male respondents and the number of female respondents (over 4 times male than female).
- Looking at the structure of respondents can help us determine whether the dataset is biased due to unbalanced distribution

Findings from Exploratory analysis – Features





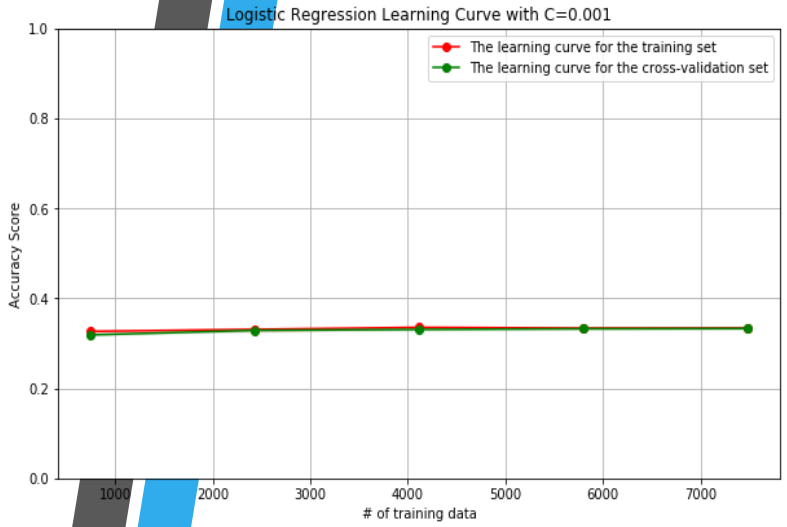
- The box plot of "Respondents' Salary Range vs. Years of Using Machine Learning Methods' shows a very interesting trend that **the more years the respondent is experienced in machine learning methods, the more average salary the respondent earns**. This plot implies that "Years of using machine learning methods" might have a strong correlation to the salary.
- The box plot of "Respondents' Salary Range vs. Age' also shows a general trend that the older the respondent, the more average salary the respondent earns. This plot implies that age might have potential correlation to the salary.

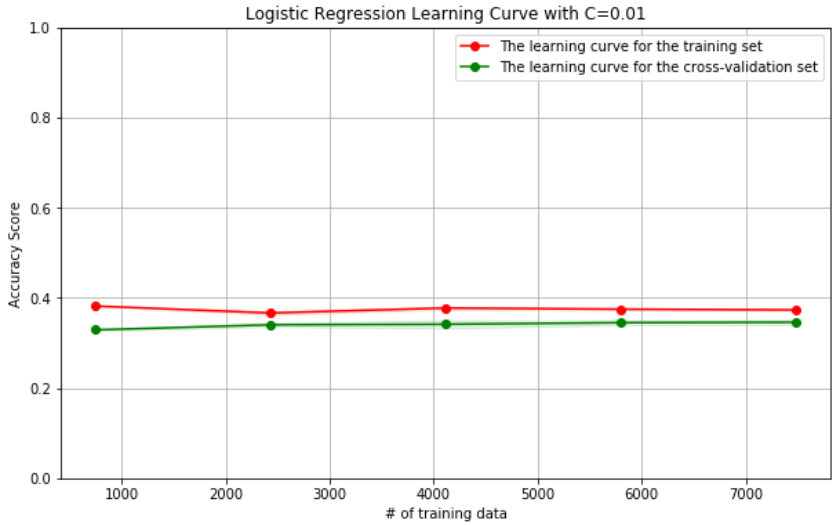
Model Feature Importance

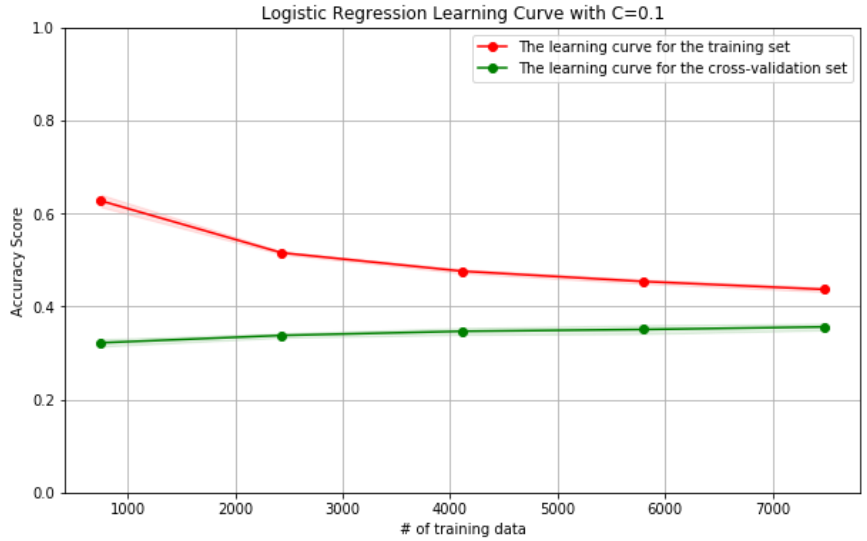
- As shown in the plot the Top 5 columns with highest correlation coefficient to the target are:

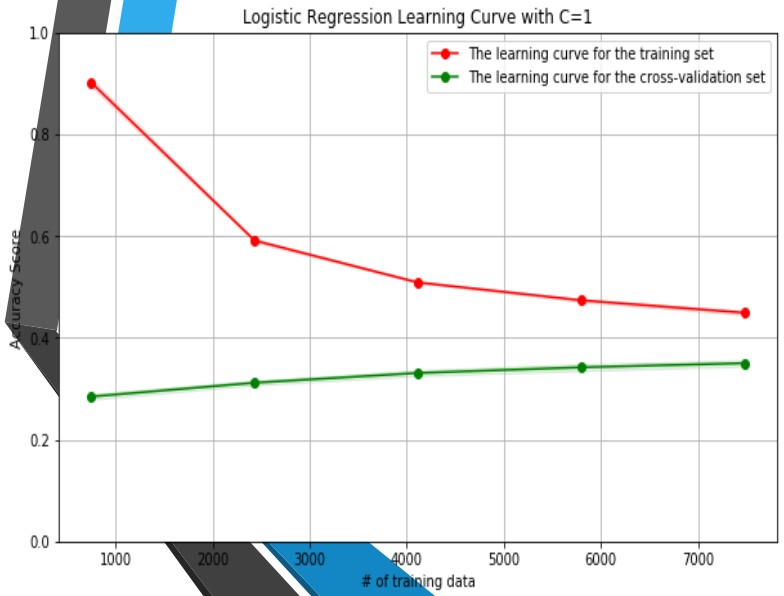
1. Q3: Whether the respondent is from USA or not
2. Q15: "How long have you been writing code to analyze data"
3. Q23: "For how many years have you used machine learning methods?"
4. Q1: "What is your age (# years)?"
5. Q11: "Approximately how much money have you spent on machine learning and/or cloud computing products at your work in the past 5 years?"

It confirmed my hypothesis in the previous slides that the respondent's salary is most likely related to the respondent's age, years of experience in /amount of money spent on machine learning and coding experience.



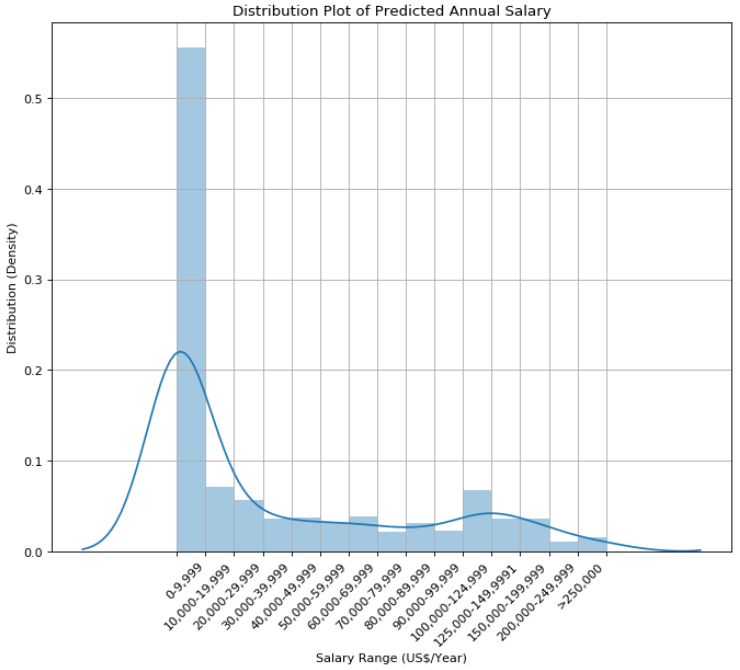


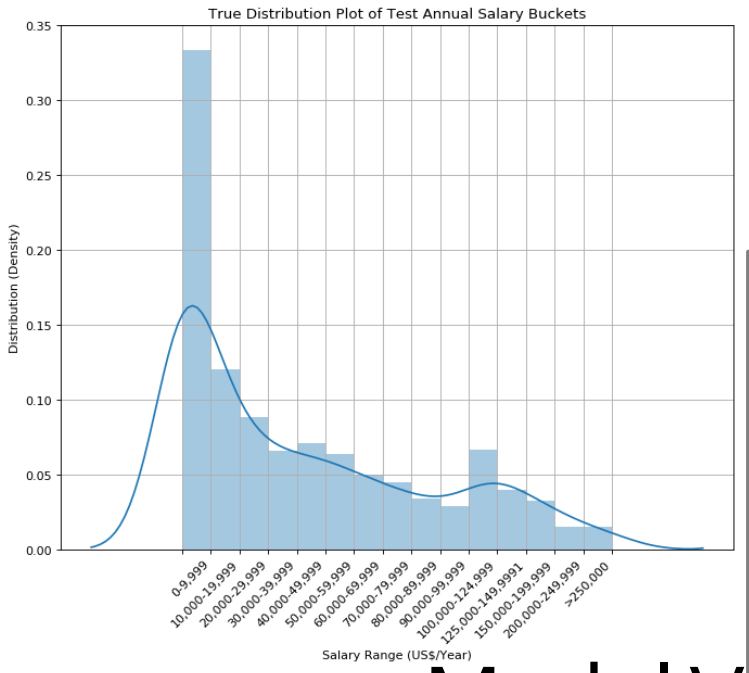


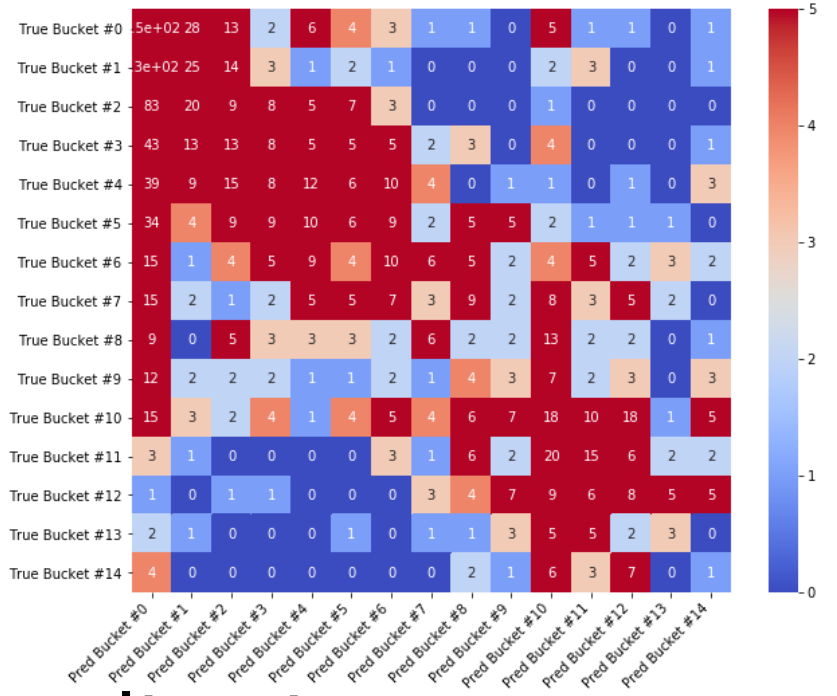


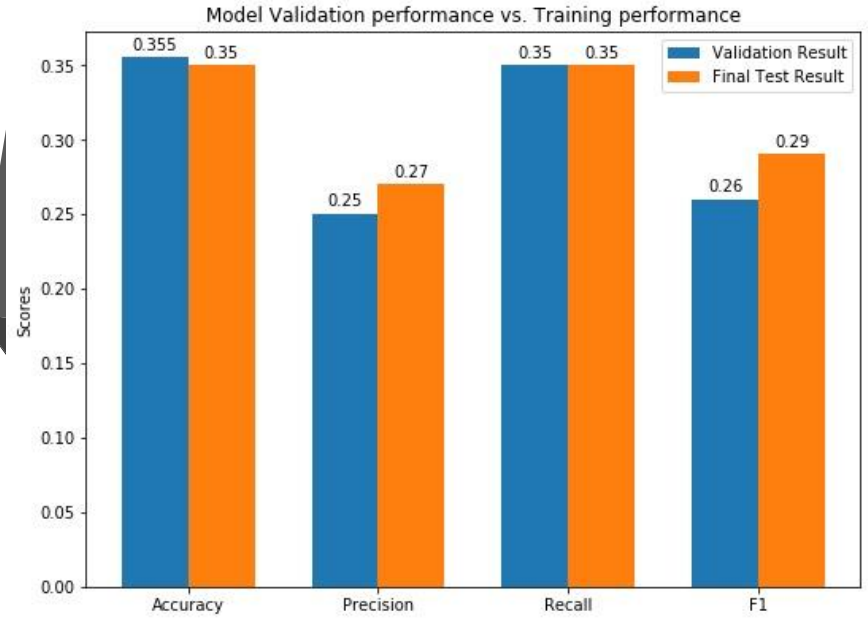
Model Training (Tuning the C value in Logistic Regression Model)

- With very lower C value (C=0.001), the training score is close to the test score and they converge. Both training accuracy and test accuracy are relatively low, this indicates that this model has high bias and low variance
- With the C value going higher and higher, the training score gradually improves and test score slightly improves, however, the gap between the training score and test score become larger and larger. This means that the model has low variance and high bias
- Thus consider the trade-off between bias and variance, the ideal C value should be somewhere in between 0.01 and 0.1 and it is confirmed from grid search that best C=0.07









Model Visualization

- The performance of the tuned best model on the validation data and test data is very close in terms of accuracy, precision, recall and f1. It proved that **the model is not overfitting**.
- By comparing with the distribution plot and looking at the heat map of the confusion matrix we can see that **the trained model is underfitting** since it has a higher distribution of bucket 0 and bucket 10. The model only captures the major trend and thus the model is **heavily biased** and underfitting