Executive Summary

In order for a business to make smarter decisions about expanding their staff, they need to understand the potential costs of bringing certain roles on to the team. By being able to forecast salaries, organizations can be more informed about the initial cost of the investment in a new employee, more informed about the cost of the investment over time, and reduce retention by anticipating employee satisfaction with their salary depending on the trends in the industry.

There were three main factors in building the model that predicted salaries: location, job title, and historical salary trends (including monthly average salaries along with year over year change). Across multiple job titles, salaries consistently increased from the beginning of when the data was tracked to the end. While the pattern at which the increase happened varied from job title to location, an increase was consistently noted.

The location variables included multiple metropolitan areas along with the national average. When comparing average salary trends with the location variable, we discovered that the salaries for these metropolitan areas were consistently above the national average no matter what the job title was being analyzed. While being above the national average was consistently demonstrated, the salary of specific job titles had ranges that were various among the regions. Additionally, it was discovered that these salaries generally

Overall, predictions for salary trends were achieved with strong accuracy. If the organization needed to forecast a salary trend for a particular job title in a particular region, the model was able to make a strong prediction in this regard.

Final Paper

Abstract

In order for a business to make smarter decisions about expanding their staff, they need to understand the potential costs of bringing certain roles on to the team. This research paper seeks to understand how factors such as geographic location, job title, and national average impact the trends that salaries take. If a business is equipped with this information, they can make smart decisions about what salary to prepare to offer a candidate that is fair for the market as well as anticipate what to budget for in the future if that role needs to continue within the company.

Intro/Background of the Problem

One of the largest expenses to an organization is the salary of its employees. However, it is operationally vital for a business to pay its employees competitively to prevent employee attrition and increase workplace satisfaction. Due to economic changes, salaries may increase in a certain field due to availability of candidates or other economic changes. In order to staff a business successfully while also planning a budget, a business needs to know what the salary trends are, where they are going, and what types of employees may be impacted.

When considering hiring a candidate, one of the most important aspects for a company to consider is how that salary will impact the profit and loss statement for that company. If it is a new role, the company needs to plan for how that salary will increase year over year. Companies need to know if a position that is filled has an expectation of a faster salary increase than other roles. It is important for a company to understand this in order to recruit the right employees and also to make sure they are prepared to address the employee's ongoing satisfaction with their salary compared to their peers in order for the company to retain the employee.

Methods

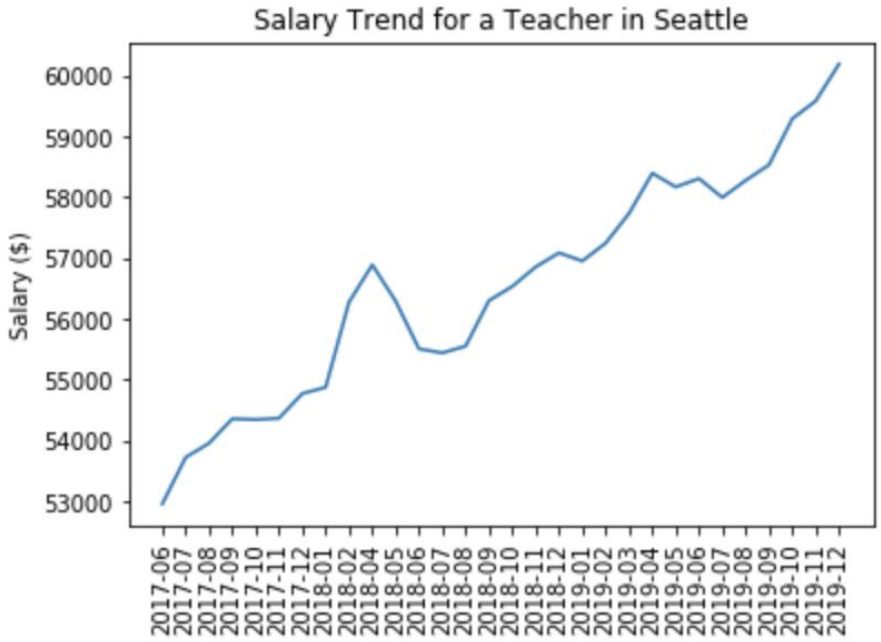
The data available is from GlassDoor's monthly reports on the job market. The months available range from June 2017 through December 2019. There is also data available for city employees of major metropolitan areas including New York City and San Francisco but they have limitations such as the date ranges available and consistency with job titles. These datasets were not included in this project but could be options for making a more complex model in the future. GlassDoor's data provides consistency with job titles to make sure the salaries and features being analyzed are for jobs that are as similar as possible.

The overall method is to analyze the historical trends of salaries in various positions in order to project how they will increase or decrease going forward. The GlassDoor datasets include a Year over Year percentage that can be studied to see if it is consistent month to month or if that metric is also changing as time goes on. Specifically, it is necessary to analyze which features are the strongest predictors for salaries, why they may be important, and their relationship with each other.

There are four main features in the dataset that have significant impact on the salary and its change over time: job title, geographic location, national average, year over year change. I started exploring each of these variables to understand their role in determining salary.

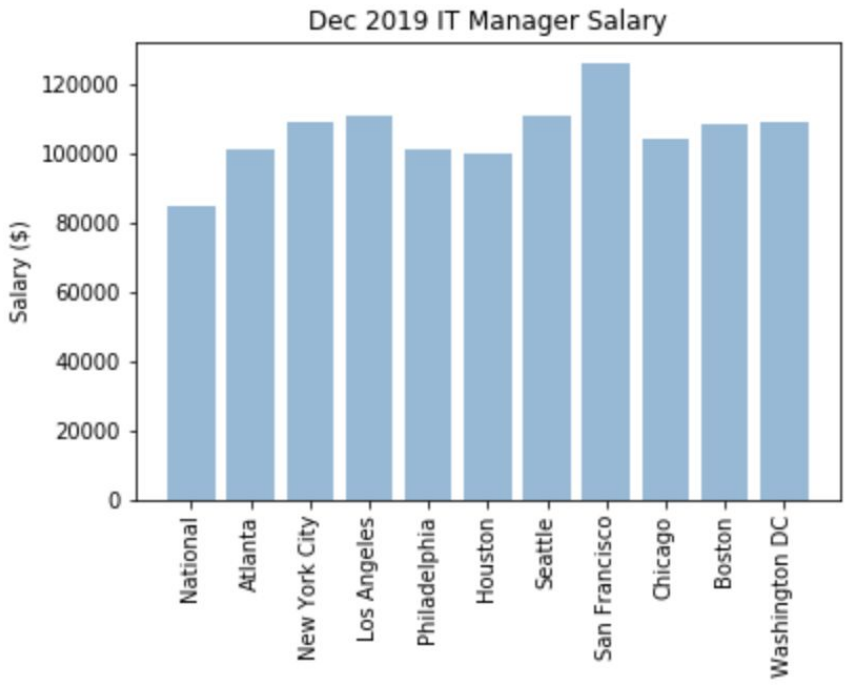
For my preliminary analysis, I focused on two different job titles to see if there were any initial insights into the trends. I chose Teacher and IT Manager as these are comparably different roles and will help with understanding any assumptions we may have. For the initial analysis, I looked at these roles in the city of Seattle. My assumption going in is that we would see an increase in salary over time.

When analyzing the salary range for a Teacher, we do see that there is an increase over time. One of the items that stands out is the bump in February through April of 2018. There is another increase in April of 2019, although smaller, which could signify a trend for this particular job, the economy, or a specific event.

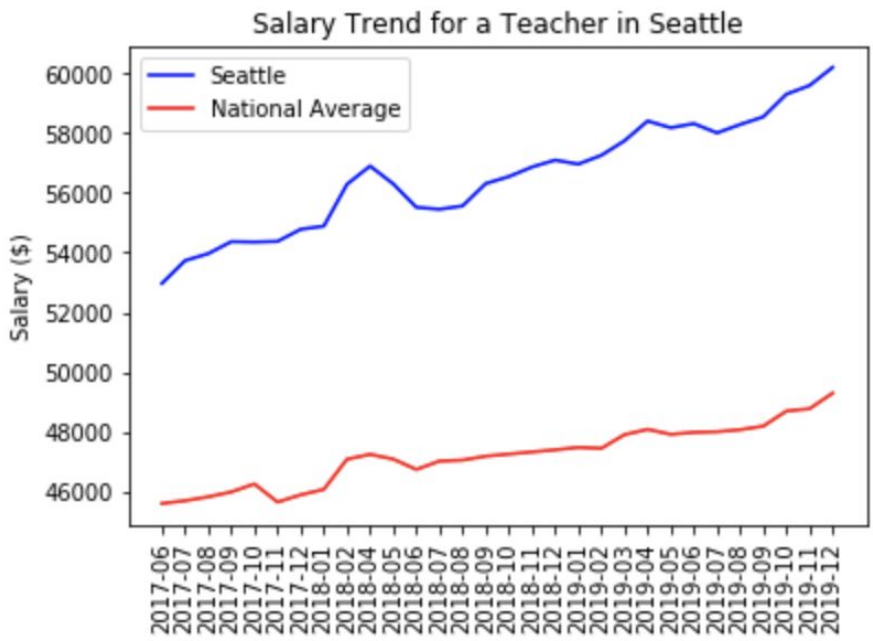


For comparison, I also looked at IT Manager for Seattle. While the overall salary does increase over time, it is much more volatile than the Teacher salary. While other aspects of the data may help in understanding these peaks and valleys, it is important to consider that outside factors could also play a part in these changes.

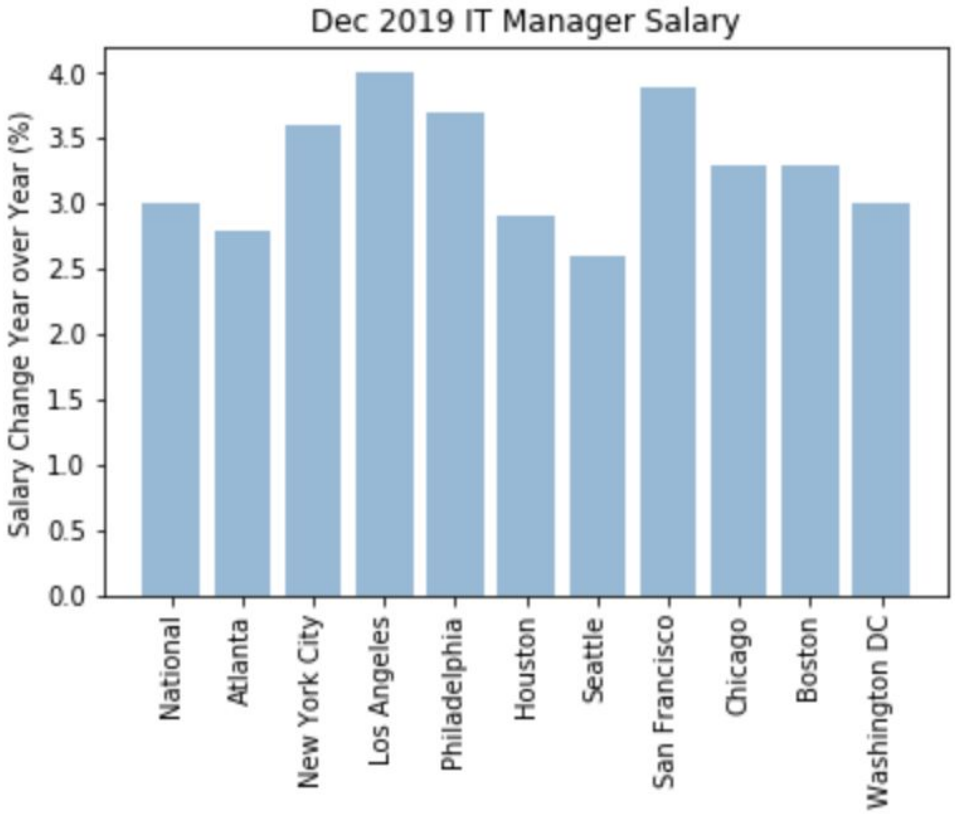
Another feature in the data that needs to be understood is geographic location. The datasets provide metrics for several U.S. metropolitan cities along with the national average. When exploring the IT manager salary across cities, we can see that there is variation. The highest salary is in San Francisco which may have been expected due to its centrality to the tech industry and famously high cost of living.



There is another insight in the previous graphic that helps us to understand the variable of the national average. The datasets being used focus on metropolitan areas but there is potential that these cities drive a higher salary demand than the national average. The trend of a metropolitan area trending higher than the national average is also visible in the Teacher salary when compared with Seattle. The y-axis was adjusted for the below graphic so that the month-to-month changes could be more easily seen. An additional insight into the Teacher salaries is that the National Average also saw a sudden spike in February through April in 2018. This signifies that the reason for salary increases at this time are largely driven by a national event or standard and not something specific to the city of Seattle.



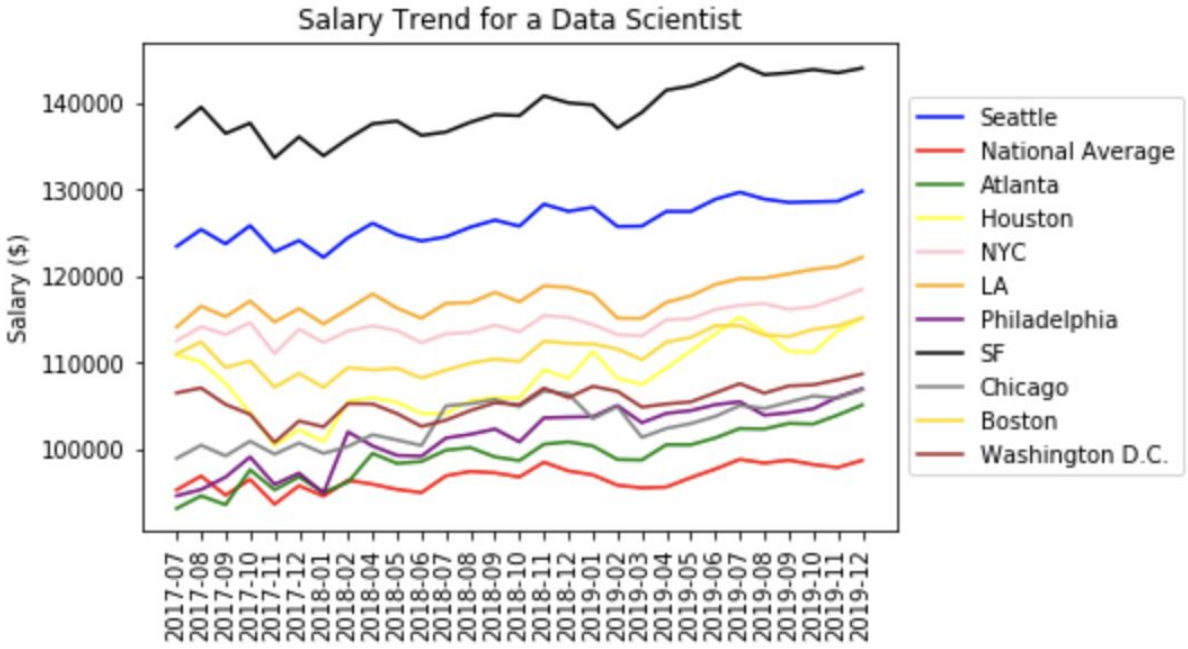
The final feature that needs to be explored is the year over year averages. This metric can potentially provide direct insight into how the salaries are changing over time and can be an important forecasting tool for where salaries will be in the future. As we explore with the December 2019 data for an IT Manager, we begin to see why it is important to examine this number instead of the salary averages for each metropolitan area. In Atlanta, Houston, and Seattle, the year over year change is increasing at a slower rate than the national average. While the salary averages are above the national average in December 2019, the slower increase rate could signal a potential future where the gap between the cities and national averages decreases.



The model building for this project was built into five phases. With each sequential phase, the model and/or the dataset became slightly more complex.

In Phase 1, the focus was on building a Decision Tree model with a binary Target Variable. The model focused only on the data for entries with the job title of Data Scientist. The Target Variable was for the salary to be either below $100 thousand or above $100 thousand.

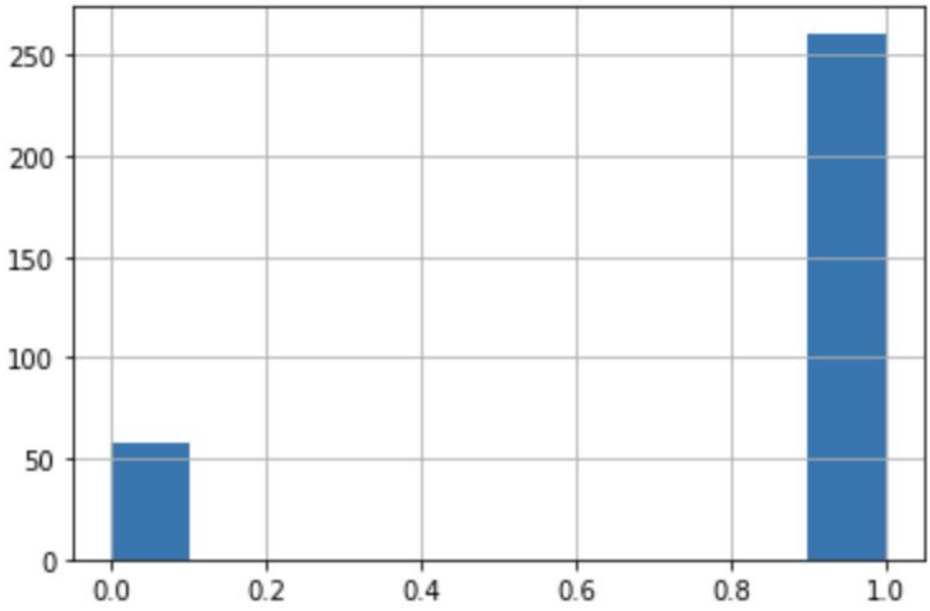
The graph below shows that the Data Scientist salary trends were a relatively normal representation of what was discovered for several roles in the dataset. The salary increased over time and the metropolitan areas were above the national average yet had their own variation in amounts.



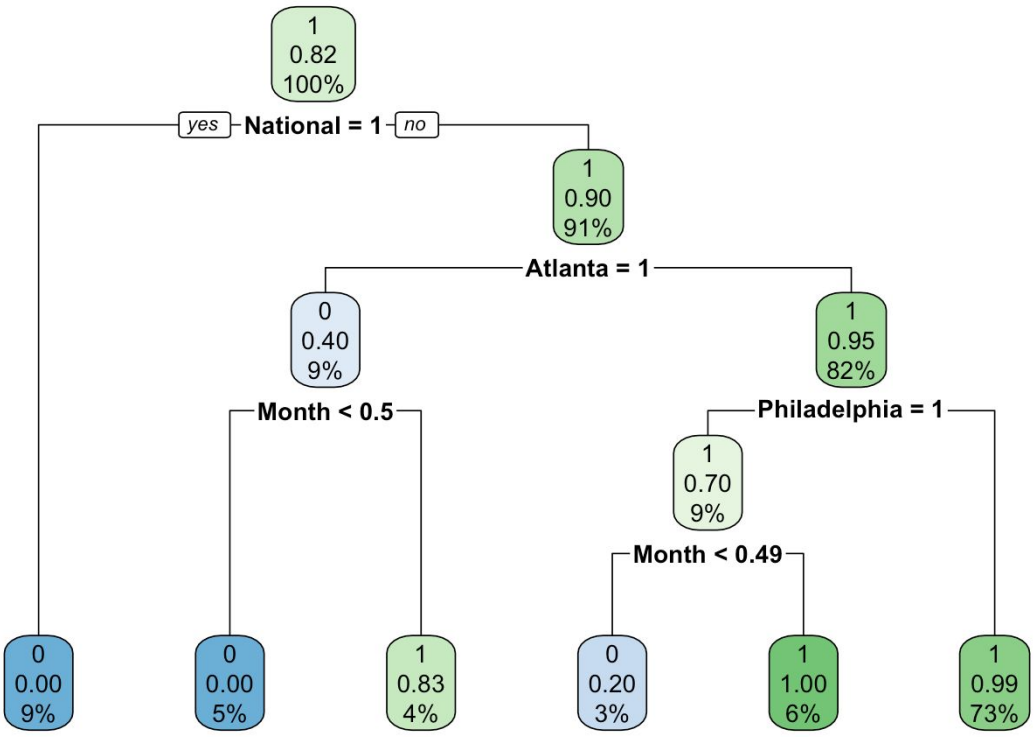
In order to use the region as a variable, one-hot encoding was used to turn each metropolitan area and the National Average into a binary variable.

As previously mentioned, the salary target variable was split to be above or below $100 thousand. The histogram below shows the distribution of the created target variable.

```
: df_Data_Scientist['Value_over100K'].hist()
: <matplotlib.axes._subplots.AxesSubplot at 0x125a5c350>
```



The image below shows the Decision Tree model that was built in R. The accuracy was 93.75%. The variables that were used for the decisions include the National Average, Atlanta, Philadelphia, and the month and year of the salary.




```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$Value_over100K,p)
accuracy <- sum(diag(confMat))/sum(confMat)
accuracy*100
```

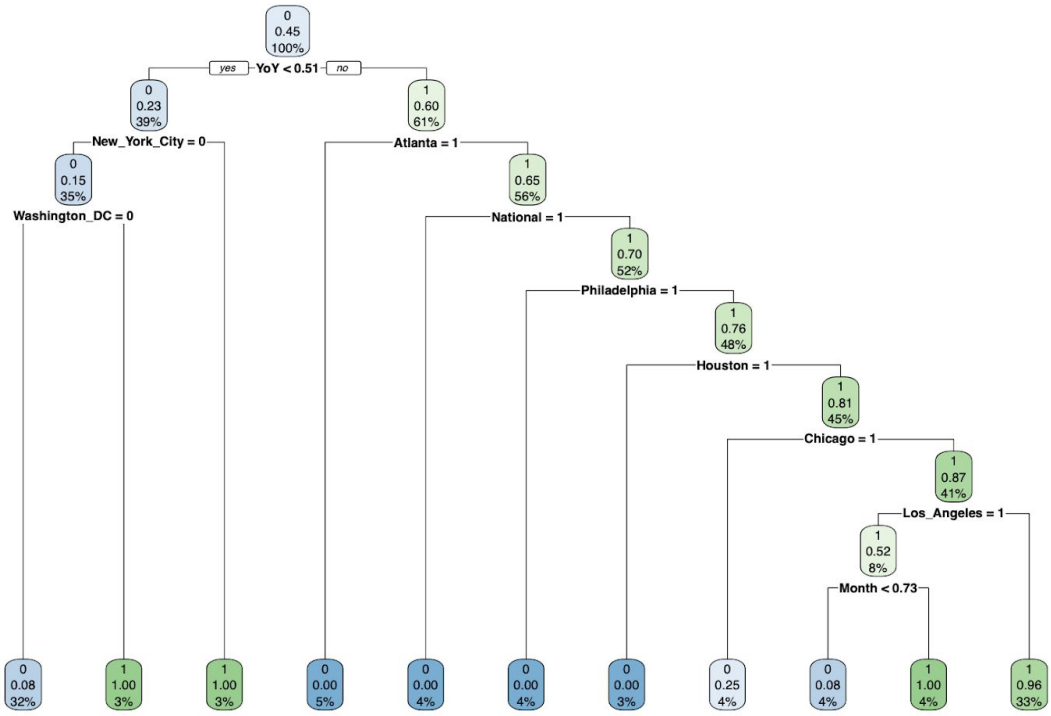
```
## [1] 93.75
```

In Phase 2, the Decision Tree model needed to be expanded to include more job titles. In order to make this more efficient, functions were created in Python to streamline the preparation of the datasets and a function was created in R for the Decision Tree model.

The target variable was also modified for these models. While $100 thousand may have been an appropriate threshold, this is not true for all job titles. In the data preparation phase for these models, the threshold was set to the mean salary for each job title. In other words, the target variable was set to above or below the average salary for the job title.

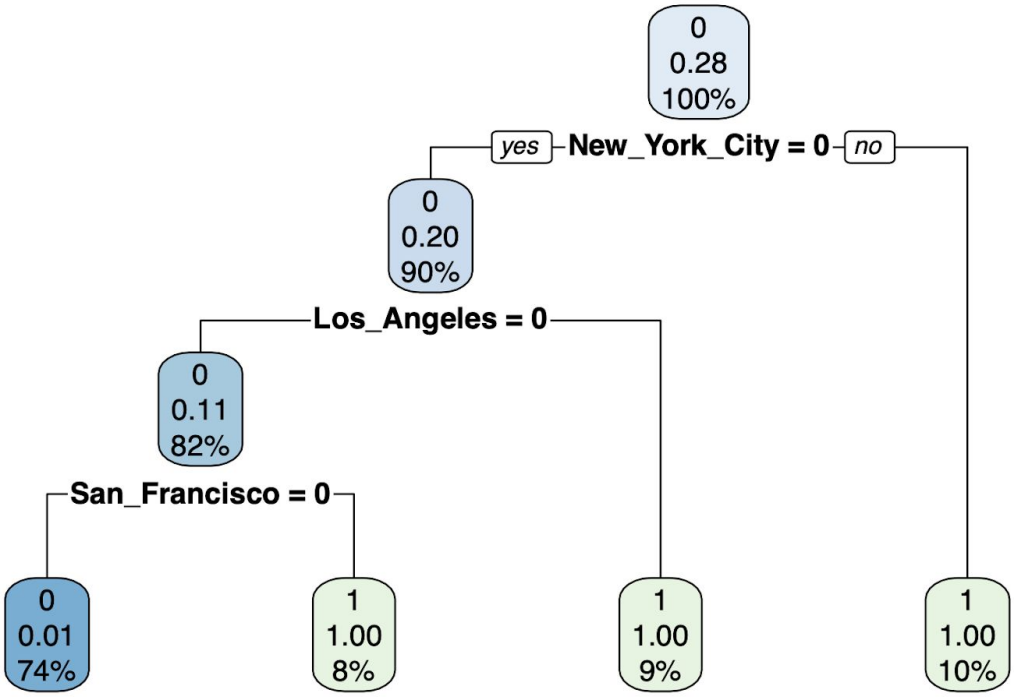
The models were consistently able to deliver results above 90%. However, the variables used (and the number of variables used) by the Decision Tree algorithm were not consistent across each model.

In the below example for Accountant, multiple variables were used to gain a 93.75% accuracy.



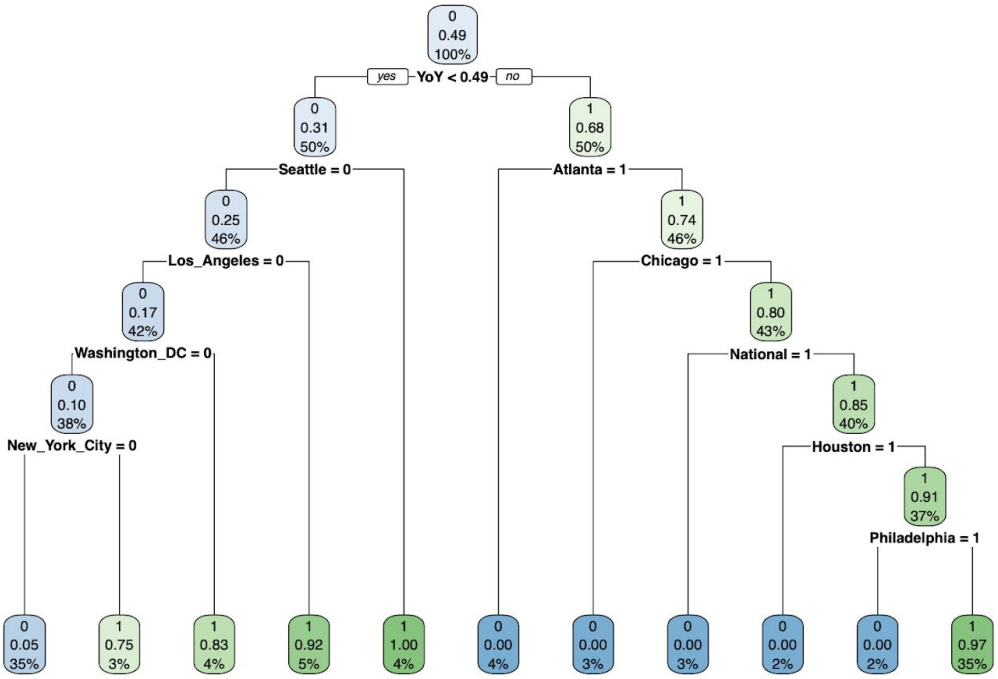

```
## [1] 93.75
```

However, the Decision Tree for a Registered Nurse used only three variables to get a 100% accuracy.

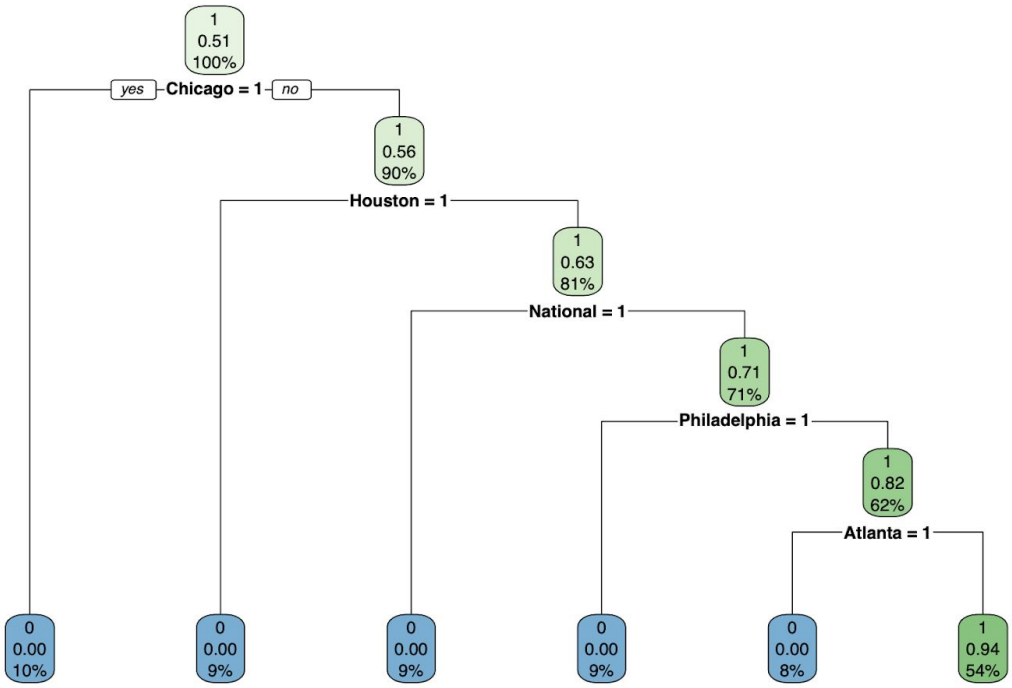



```
## [1] 100
```

A couple additional examples are below of some of the Decision Trees that were built. The first is for Data Analyst (with 100% accuracy) and the second is for Graphic Designer (with 90.625% accuracy).


0
0.49
100%
yes
YoY < 0.49
no
0
0.31
50%
1
0.68
50%
Seattle = 0
Atlanta = 1
0
0.25
46%
1
0.74
46%
Los_Angeles = 0
Chicago = 1
0
0.17
42%
1
0.80
43%
Washington_DC = 0
National = 1
0
0.10
38%
1
0.85
40%
New_York_City = 0
Houston = 1
1
0.91
37%
Philadelphia = 1
0
0.05
35%
1
0.75
3%
1
0.83
4%
1
0.92
5%
1
1.00
4%
0
0.00
4%
0
0.00
3%
0
0.00
3%
0
0.00
2%
0
0.00
2%
1
0.97
35%

```
## [1] 100
```




```
## [1] 90.625
```

In Phase 3, it was important to address the major limitation in the previous two phases which is the target variable. Predicting whether a salary is above or below a threshold is not necessarily informative to an organization trying to have a comprehensive understanding of salary trends.

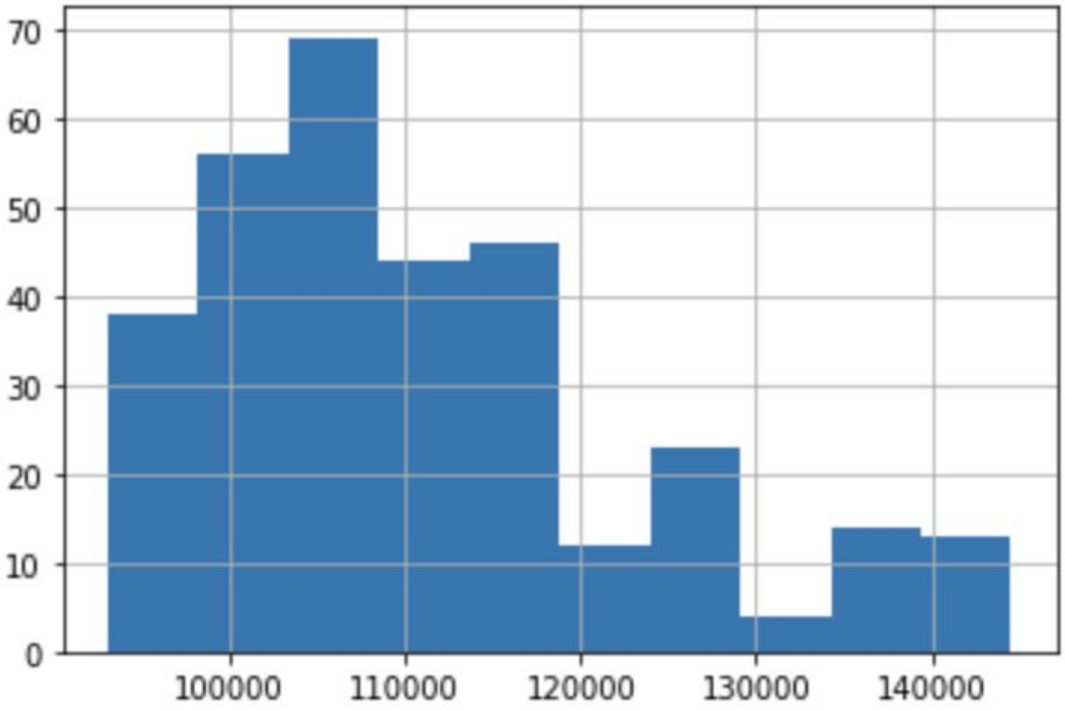
The target variable needed to be placed in bins and then run through the Decision Tree model with these new targets. Phase 3 returned focus to the Data Scientist dataset. The target variable was split into the following five bins:

1) Below $100 thousand
2) $100 thousand to $110 thousand
3) $110 thousand to $120 thousand

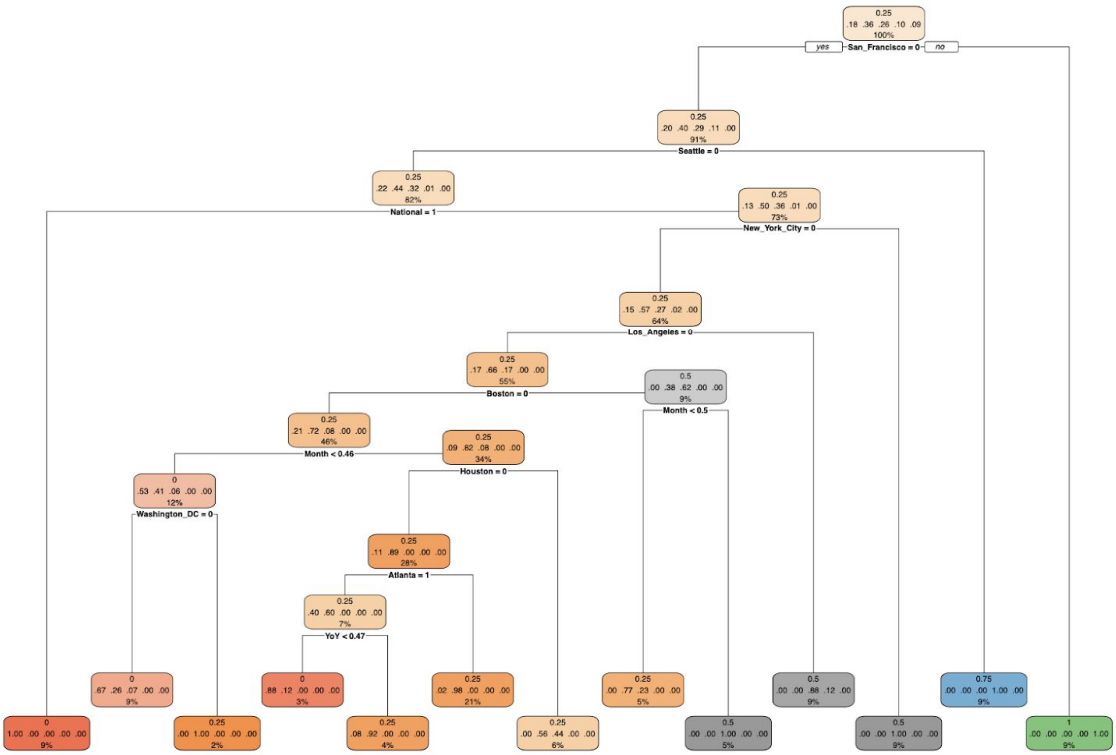
4) $120 thousand to $130 thousand

5) Above $130 thousand

Here is a histogram showing the distribution of salaries for Data Scientist which led to the decision of how the bins should be set.



The Decision Tree was still effective but the accuracy did decrease to 81.25%. The model is illustrated below.



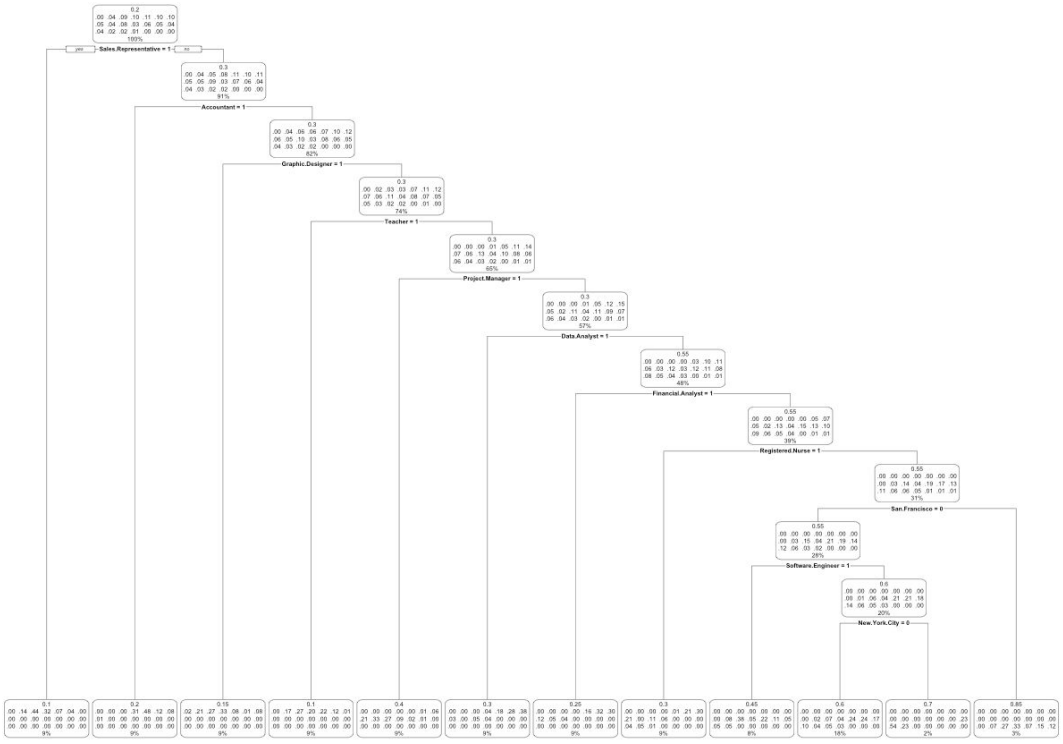

```
# [1] 81.25
```

Since the Decision Tree still worked in Phase 3, it was important to continue using this algorithm in Phase 4. However, there were two changes that needed to be made. The bins set for the target variable in Phase 3 were not going to work for all job titles. The first change needed to expand the target variable to include enough bins for all salaries. 28 bins were created for the target variable for the next model. They increased in value every $5,000 starting with $20,00 and ending with $150,000. There was also a bin for less than $20,000 and another bin for above $150,000.

The second modification was to make the job title a variable. The Glassdoor dataset includes over one hundred job titles. However, for this model, only twelve were selected. The

twelve job titles used were Software Engineer, Project Manager, Financial Analyst, Accountant, Sales Representative, Professor, Registered Nurse, Teacher, Graphic Designer, Data Scientist, Data Analyst, and IT Manager. These values were then one-hot encoded so that each row had a binary value for each of the selected job titles.

The Decision Tree was not successful with this dataset and only achieved 14.63%. The reasons for this are quite clear. Decision Trees work best with a limited number of target variables. A binary target variable such as the one in Phase 1 and Phase 2 were suited for a Decision Tree whereas 28 is far too many for it to be successful.




```
## [1] 14.63415
```

In Phase 5, the decision was made to use the same dataset as in Phase 4 but with a different algorithm. A neural network was built in Python. The neural network used the same dataset and was able to achieve an accuracy of 85.48%. The details of the results and the confusion matrix are included below.

	precision	recall	f1-score	support
5.0	1.00	1.00	1.00	1
6.0	0.94	0.97	0.95	31
7.0	0.93	0.91	0.92	57
8.0	0.87	0.90	0.88	72
9.0	0.92	0.85	0.88	78
10.0	0.84	0.95	0.89	62
11.0	0.91	0.90	0.90	89
12.0	0.90	0.83	0.86	46
13.0	0.84	0.76	0.80	34
14.0	0.82	0.92	0.87	65
15.0	0.73	0.50	0.59	16
16.0	0.84	0.80	0.82	51
17.0	0.76	0.79	0.77	39
18.0	0.81	0.74	0.77	23
19.0	0.77	0.86	0.81	28
20.0	0.76	0.81	0.79	16
21.0	0.46	0.75	0.57	8
22.0	0.89	0.53	0.67	15
23.0	0.00	0.00	0.00	1
24.0	1.00	1.00	1.00	4
25.0	1.00	1.00	1.00	1
accuracy			0.85	737
macro avg	0.81	0.80	0.80	737
weighted avg	0.86	0.85	0.85	737

```
[[ 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 30  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  2 52  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  3 65  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  7 66  5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  2 59  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  6 80  3  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  7 38  1  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1 26  7  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  4 60  1  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  6  8  2  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  2 41  8  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  6 31  2  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  2 17  4  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  2 24  2  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  3 13  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  6  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  7  8  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1]]
```

```
Model Accuracy: 85.48%
```

Results

There are two important results from the five phases of model building. The first is that Decision Tree model that does not use job title as a variable, can deliver a high accuracy when predicting the salary for a month and for a particular job title in a specific metropolitan area. The results were consistently over 90% accuracy. The second important result is that a neural network can also deliver a high accuracy if the job title is a variable that is included in the model. These two different results are highlighted as they would be the basis for the decision on next steps. The decision would revolve around building individual decision trees for each job title or to build one large neural network that takes in all of the data.

Discussion/Conclusion

The overall conclusion is that highly accurate predictions can be made for salaries at various times for different job titles in different regions. An organization could use the predictive

model to make informed forecasts on the financial impact of adding a role to the organization or filling a vacancy.

However, there are improvements that can be done to the model to make sure the results are as robust as possible. First of all, the final model only worked with twelve job titles while the original datasets included more than one hundred. In order to push the test of the neural network further, increasing the number of job titles is a possible step. However, the model used in Phase 3, which did not use job title as a variable, may be sufficient for providing the information that an organization may need for forecasting salary expectations.

Glassdoor releases salary information on a monthly basis. In order for the model to be optimized, an automated process of downloading, converting, and adding the new data to the model could improve accuracy over time. The monthly data could also provide additional insights into when salary trends are more likely to increase or decrease as well as how metropolitan areas are impacted. Additionally, there are datasets for government employees available that may be able to provide additional data points including years of experience or previous job titles.

Finally, another development would be accessibility. For optimal use, the model would need to be available to users with no technical expertise. The next steps would include a way for the model to receive inputs and then provide a forecast to the user.

References

https://www.glassdoor.com/research/job-market-report-historical/