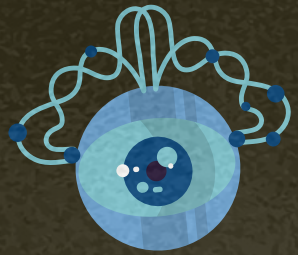
SC1015: Mini Project

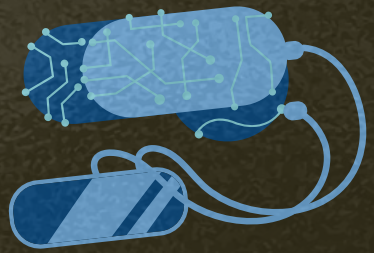
Salary Prediction

SC1 Shawn Lim (U2121495D),
SC1 Chen Yiming (U2120135L)

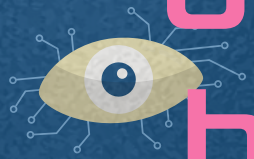


Introduction

According to the NTU official website, our Computer Science degree has the Second Highest Starting Salary in the school.



What factors and skill sets will allow us to command higher salary?



Data

Job posting and salaries taken from California

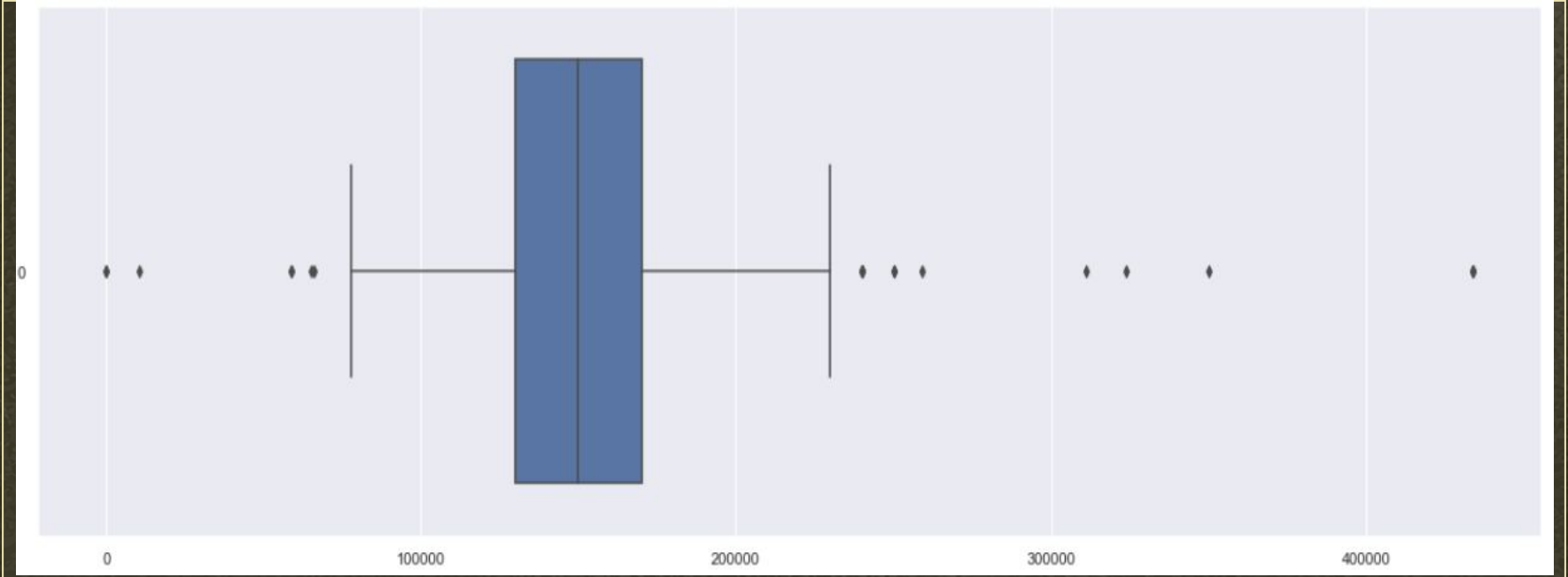
Variables

City, Company, Title, Salary and other skill requirements

Exploratory Data Analysis

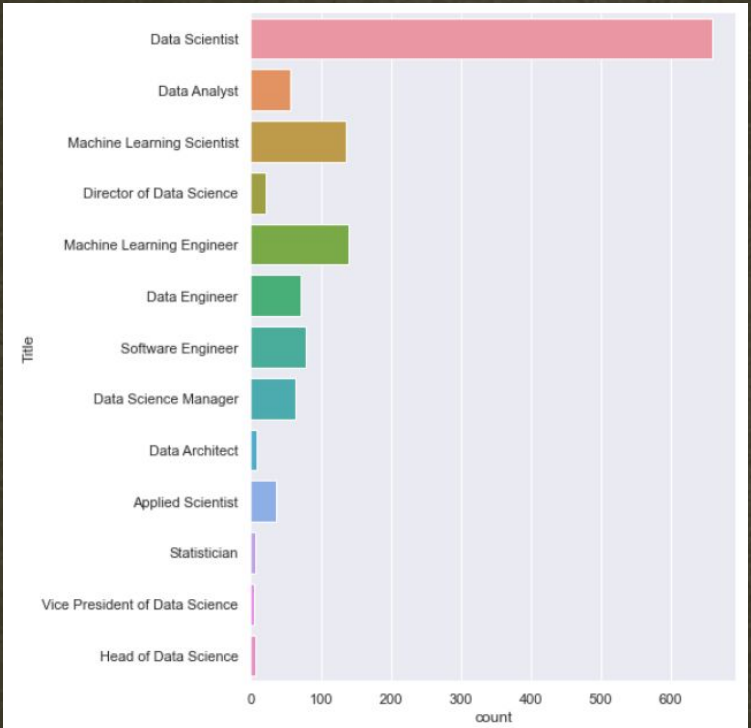
Visualising The Data!

Spread of Yearly Salary



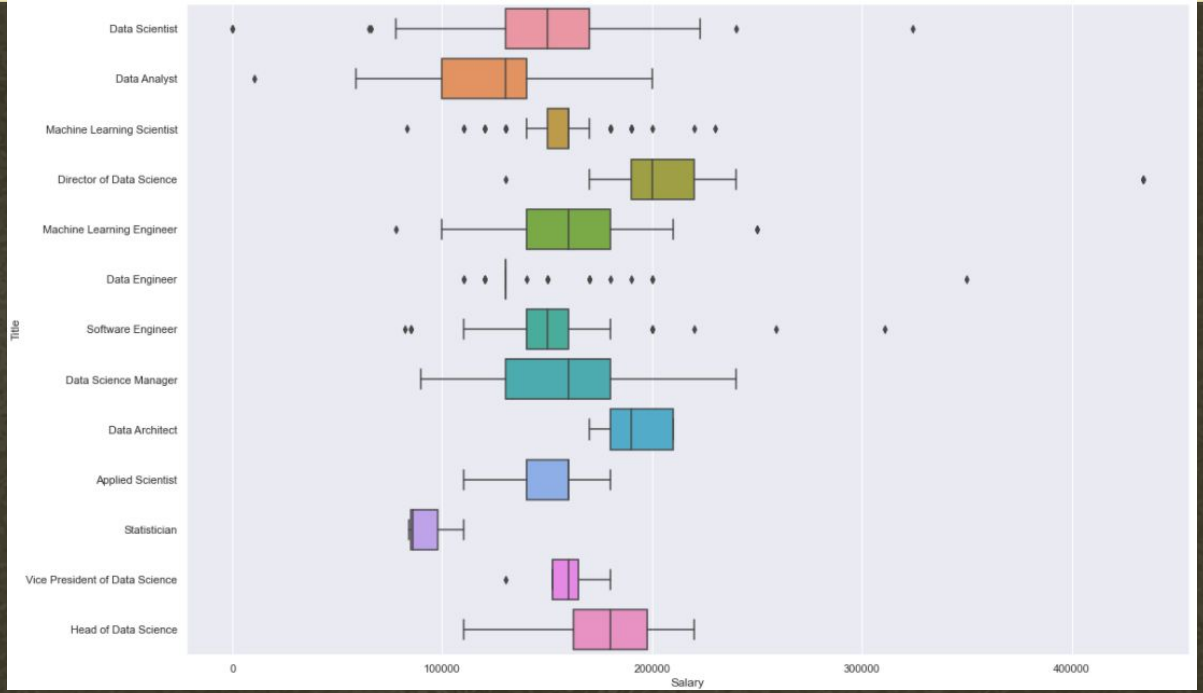
Initially it looks like yearly salary is very high of around 6 figures per annum!

Different Titles



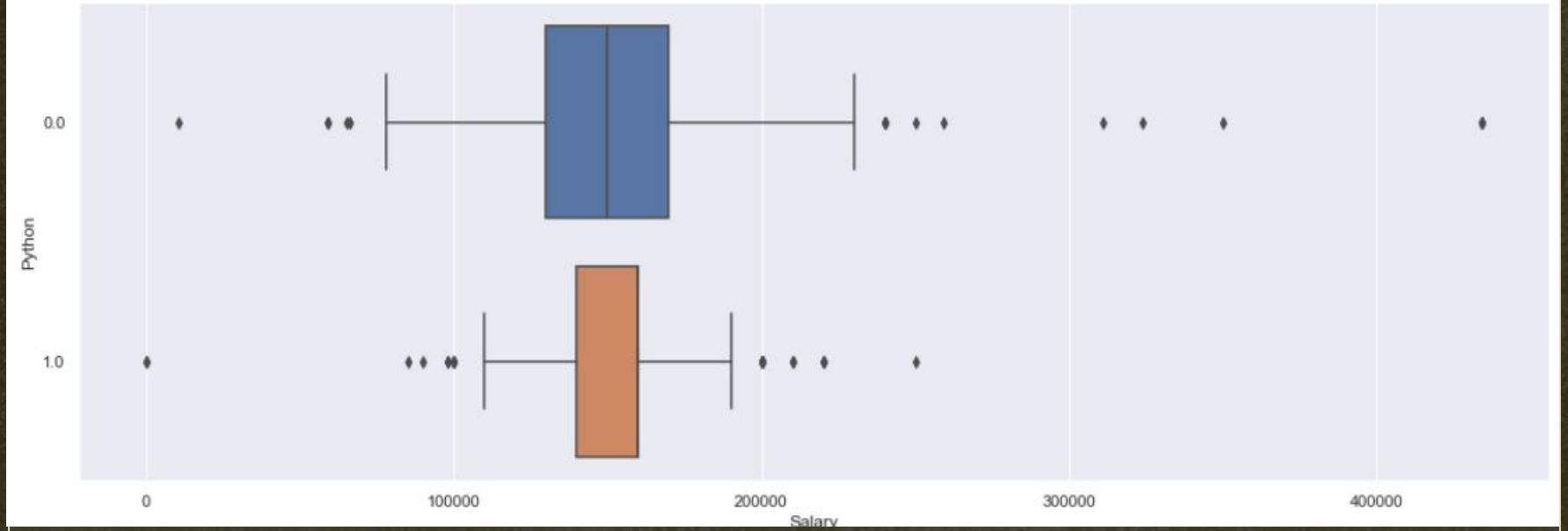
Data Set has many different jobs with around half the data set being Data Scientist

Different Titles vs Salaries



Breakdown of salaries show that the median salaries of data scientist is mostly lower

Python's Importance



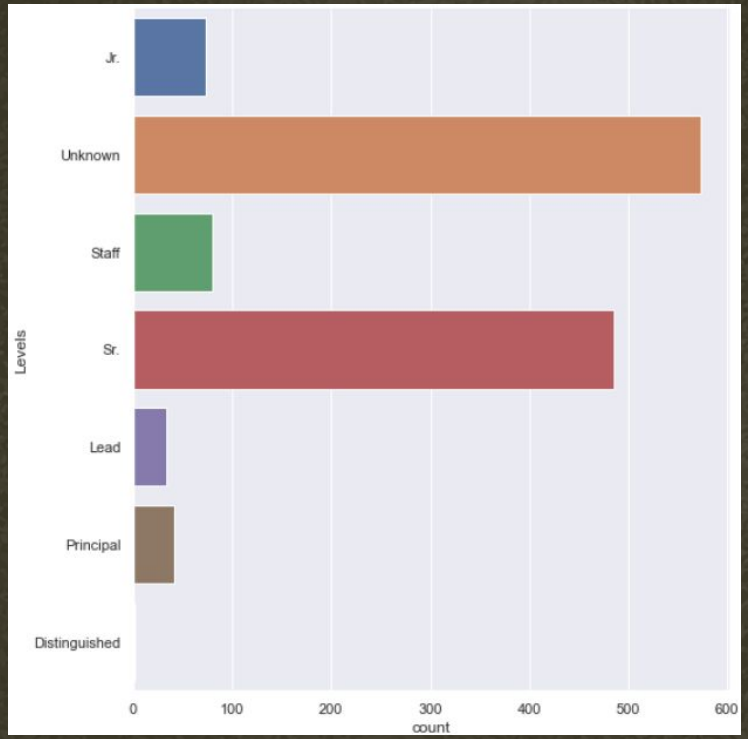
Python doesn't seem like an important determining factor

At first look it is **Impossible** to predict salary. Is there a better way to visualise the data?

Data Preparation and Cleaning

- Removing Duplicate Columns
- Deleting Unhelpful Columns
- Removing Rows With NULL Values
- Removing Outliers
- Performing Encoding

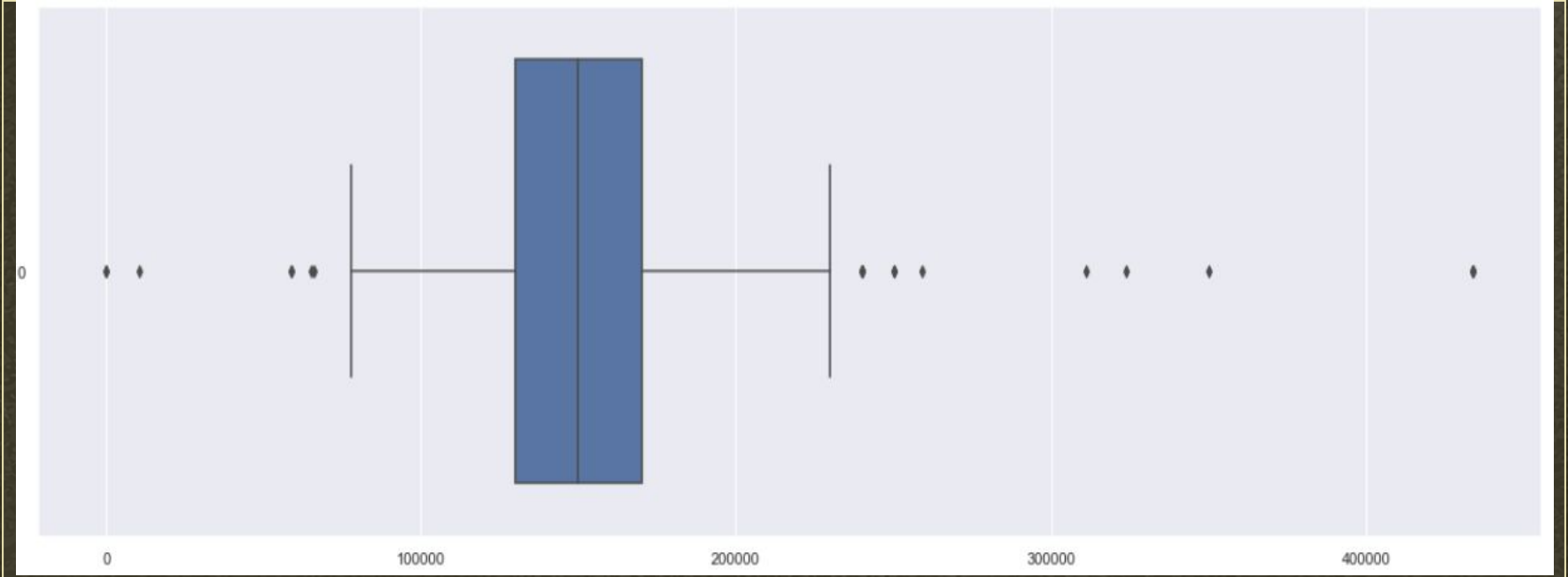
Duplicate and Unhelpful Columns



Job_ID appears for 3 times while in the levels column the number of unknown values was so large it would not be best that we do not use it

	Job_ID	Job_ID
0	0	0
1	1	1
2	2	2
3	3	3
4	4	4

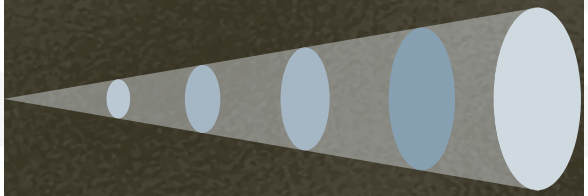
Null Values and Outliers



There were some insanely high and low salaries including 0 and Null values that had to be removed

Performing Encoding

Title	Company	City
Data Scientist	Numerdox	Sacramento
Data Analyst	Cepheid	Lodi
Data Scientist	Cepheid	Sunnyvale
Data Scientist	Verana Health	San Francisco
Data Scientist	Tinder	San Francisco



Title_encoded	City_encoded	Company_encoded
5	48	239
1	27	71
5	64	71
5	52	363
5	52	339

In order to assist in machine learning, categorical data should be transformed into numerical data.

We could either perform label encoding or one-hot encoding

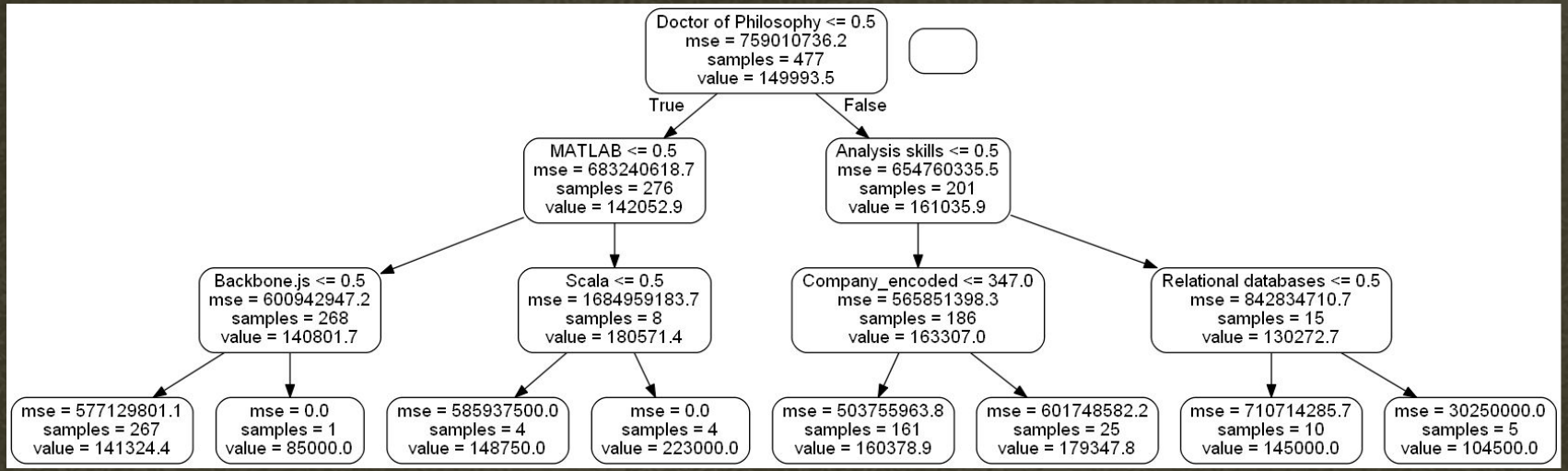
Machine Learning

Random Forest Regressor

Random Forest

1. We split the data randomly into the train and test sets
2. Import random forest regression model, instantiate the model, and fit the model on the training data
3. Make predictions on the test set

Visualizing a Decision Tree



The full Decision Tree would be too difficult to visualise so we inspect a smaller portion

Performance of Model

```
Mean Absolute Error: 14467.68 dollars.
Accuracy: 90.27 %.
```

- We use mean absolute error to see how far our average prediction differs from the actual value
- To get accuracy we take mean absolute performance error subtracted from 100%
- We have a good accuracy at 90.27%

Factor Importance

```
Variable: Company_encoded       Importance: 0.12
Variable: Doctor of Philosophy  Importance: 0.07
Variable: City_encoded          Importance: 0.07
Variable: Title_encoded         Importance: 0.06
Variable: Doctoral degree       Importance: 0.05
Variable: MATLAB                Importance: 0.04
Variable: Analysis skills       Importance: 0.02
Variable: Bachelor's degree     Importance: 0.02
Variable: Relational databases  Importance: 0.02
Variable: Azure                 Importance: 0.02
```

Seems like Company is the most important variable when predicting the salary with having a doctorate in philosophy playing an important role as well.

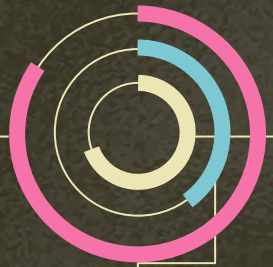
Interesting Insight

```
Variable: Company_encoded       Importance: 0.12
Variable: Doctor of Philosophy  Importance: 0.07
```

```
Mean Absolute Error: 18261.79 dollars.
Accuracy: 87.62 %.
```

When simply using the 2 most important variables, “Company_encoded” and “Doctor of Philosophy”, the accuracy barely decreases

Learning Points



Filter and sort data using label encoding

Random Forest Regression

Just having two most important variables could give a good prediction

Conclusion

- We are now able to determine what skill sets and factors could give us higher salaries
- Whichever company we work at is even more important than what we specialise in
- Maybe we should get Doctorate in Philosophy!

THANKS!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Please keep this slide as attribution

References

1. https://www.ntu.edu.sg/scse/news-events/news/detail/the-highest-starting-salary-degrees-is-at-ntu
2. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
4. https://pbpython.com/categorical-encoding.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html