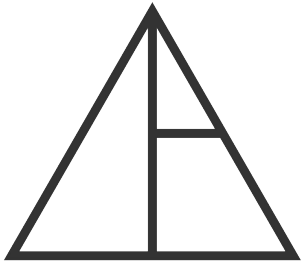


VERTICAL
INSTITUTE

Mick Yang

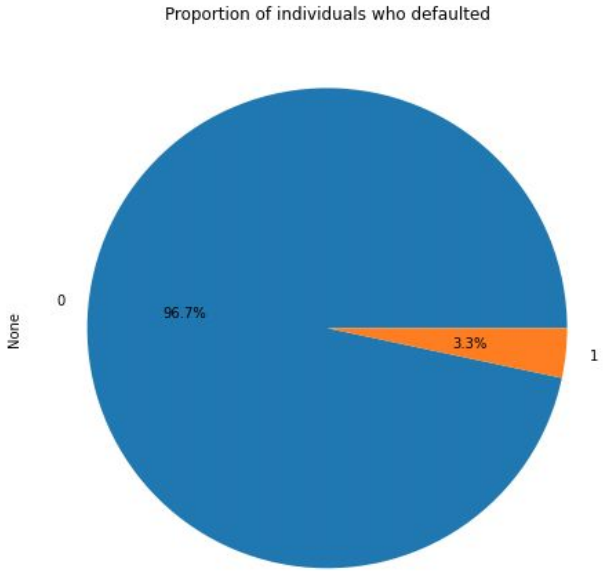
Problem

What factors matter most in predicting whether someone will default on a loan?

Possible factors

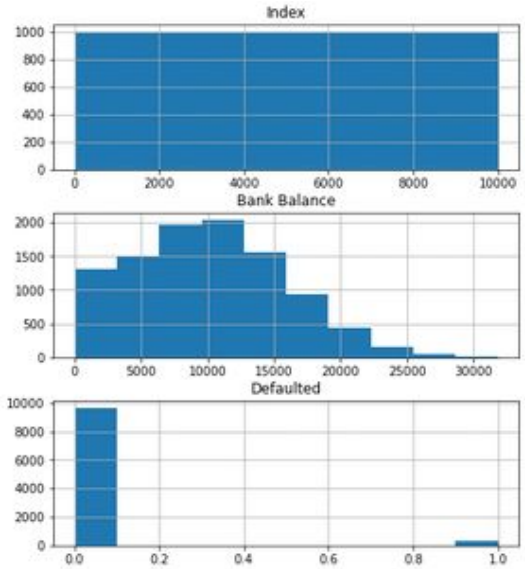
1. Employment status
2. Bank balance
3. Annual salary

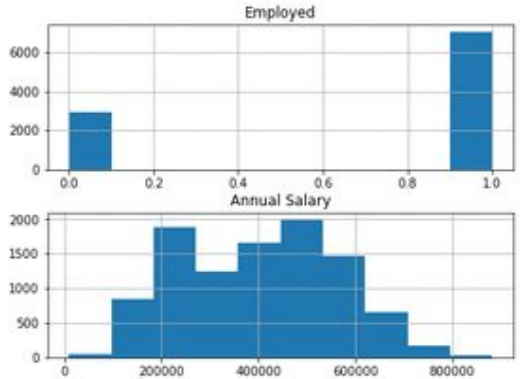
Data set exploration!



In a synthetic data set of 10,000 cases, only a tiny fraction (3.3%) result in default status

Data set exploration!

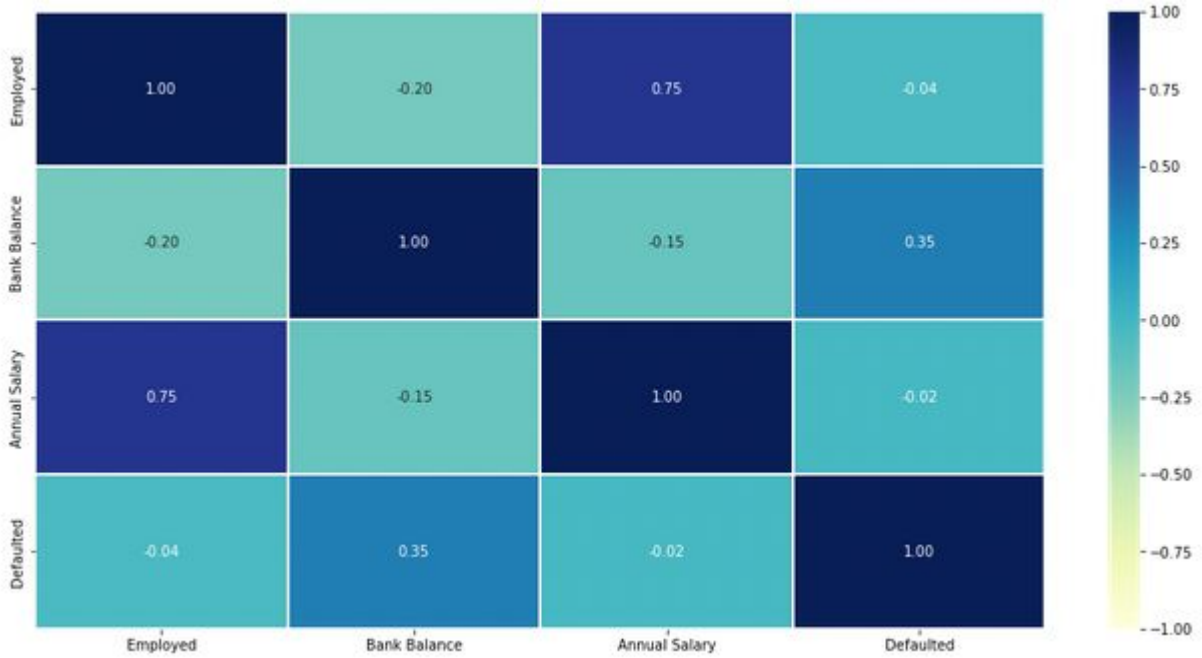




Understanding the data types through a histogram plot:

- **Bank Balance** and **Annual Salary** data are continuous
- **Employed** and **Defaulted** data are binary and categorical

Data set exploration!



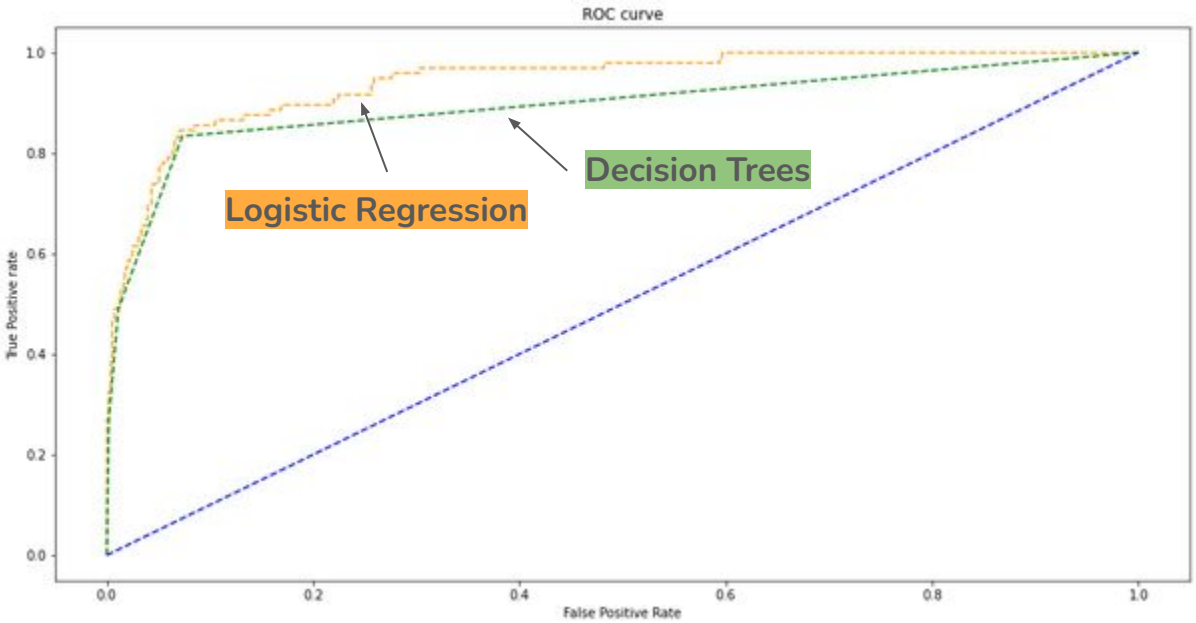
Plotting a correlation graph indicates some correlation in these relationships:

- Bank Balance-Defaulted
- Annual Salary-Employed

Choosing our model

Logistic Regression	Decision Trees
Rationale - Single linear decision boundary in feature space - Less susceptible to overfitting - Robust to noise	**Rationale** - Non-linear decision boundary - Easier to interpret - Better at handling skewed data, outliers, or missing values
Outcome - Accuracy score: 0.9612 - After standardising data, it improved to 0.9712	**Outcome** - Accuracy score: 0.9541 - After improving data with GridSearch, it improved to 0.973

Training our models



Plotting the improved **Logistic Regression** and **Decision Trees** model and it seems like Logistic Regression classification model does better

Conclusion

Findings

- Strongest correlation lay between Bank Balance and Employment Status
- No variable was particularly indicative of Default Status

Next steps

- Build a more robust dataset with more interesting variables
- Employ feature scaling of dataset
- Improve the models by tuning hyperparameters

Thank you! Any questions?

Credits

Vertical Institute bootcamp team
Sifat, *Calven*, *Safkat*, *Jacob*

Dataset taken from:
https://www.kaggle.com/kmldas/loan-default-prediction