Customer Salary Prediction

COLLIN GUIDRY

Objectives

- Develop a process to decide which customers should be targeted for our marketing campaign.
- Use census data to train a model that can predict whether an individual makes greater than $50,000 a year
- Verify the model's accuracy for predicting the income of potential new customers

Executive Summary

- Salary can accurately be predicted based on our analysis
- We chose to use a Random Forest Classifier as our tool for prediction
  - Highest accuracy rate
  - Least likely to produce costly predictions

Data Overview

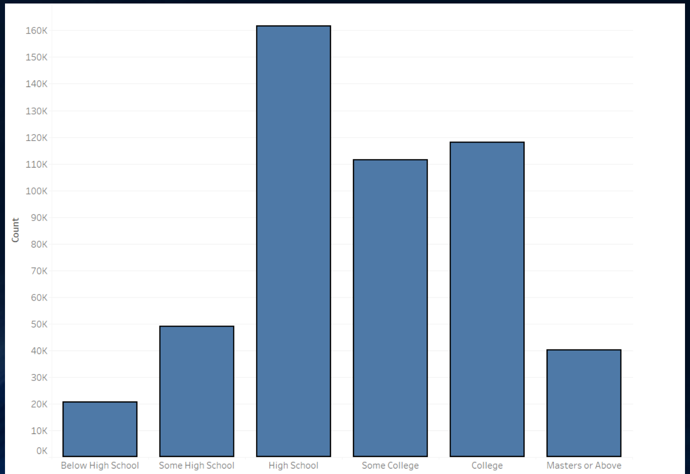
How did we sample the data that was given?

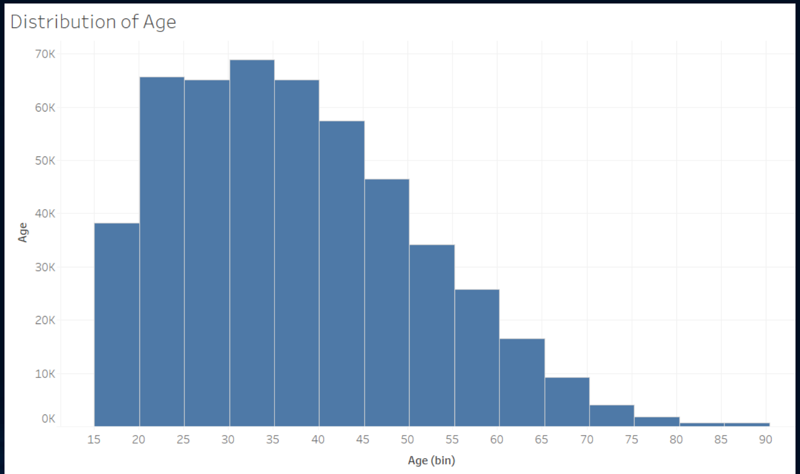
- Weight-based sampling was used to generate a dataset from census data.
  - Sampling enabled us to reduce the data to a feasible size
- Removed outliers
- Simplified variables such as education level and marital status

Data Overview (cont'd)

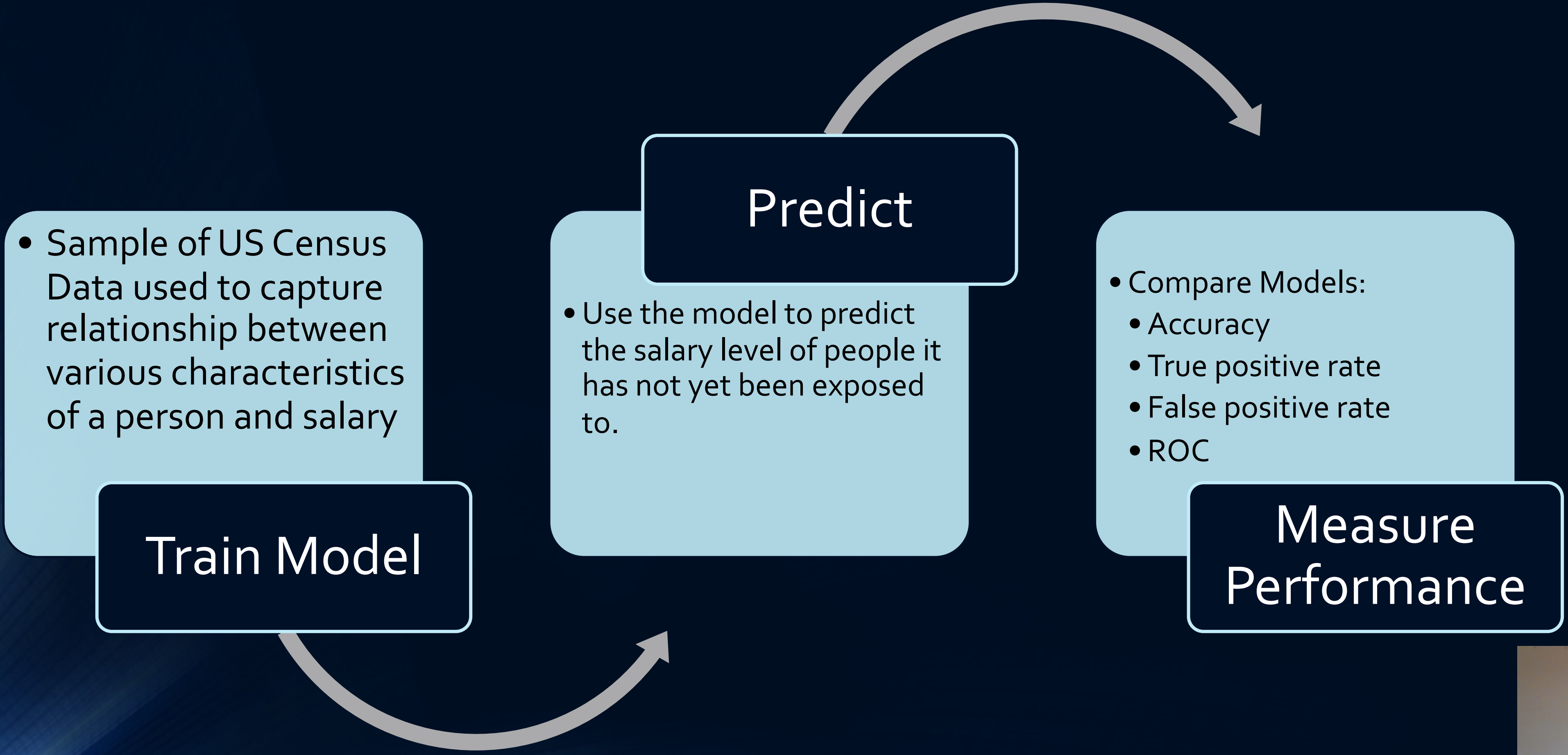
What does our sample data look like?

- 77% of people do not make over $50k
- Approximately 55% of people have either a High School, Some College, or College Education





Model Development Process



Train Model

- Sample of US Census Data used to capture relationship between various characteristics of a person and salary

Predict

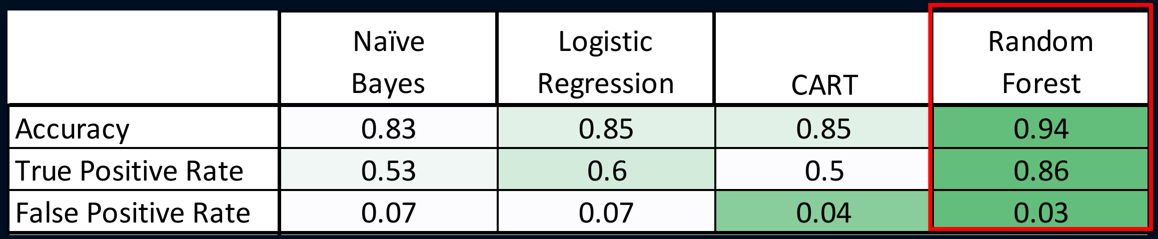
- Use the model to predict the salary level of people it has not yet been exposed to.

Measure Performance

- Compare Models:
  - Accuracy
  - True positive rate
  - False positive rate
  - ROC

Model Selection

- Random Forest Classifier
  - Most **accurate** predictions (accuracy of **94**%)
  - 86% of all correct predictions were for high income individuals.
  - Of the correct predictions, most were in our target market of high-income individuals (86%)
  - This model is the least likely to classify a low-salary person as high salary. (Most costly error)



	Naïve Bayes	Logistic Regression	CART	Random Forest
Accuracy	0.83	0.85	0.85	0.94
True Positive Rate	0.53	0.6	0.5	0.86
False Positive Rate	0.07	0.07	0.04	0.03

Refer to Appendix A & B to see further comparison of each model type

Best Indicators of High Salary

- **Capital Gain / Capital Loss**
  - Those who invest have higher salaries are more likely to have disposable income to invest

- **Education**
  - Those who reached graduate school or above are 43% more likely to make more than $50K

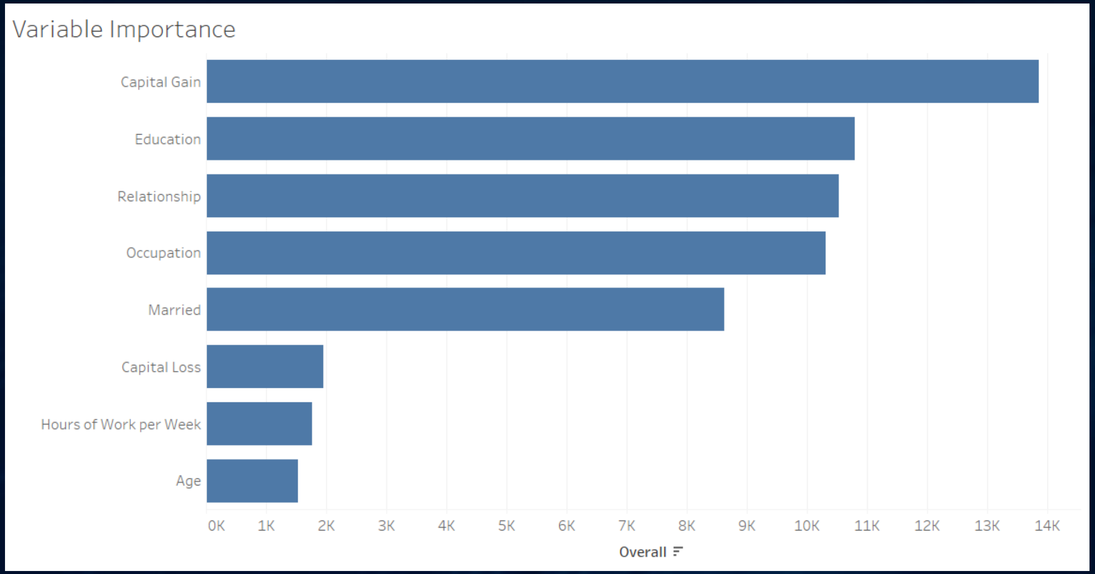


Figure 1

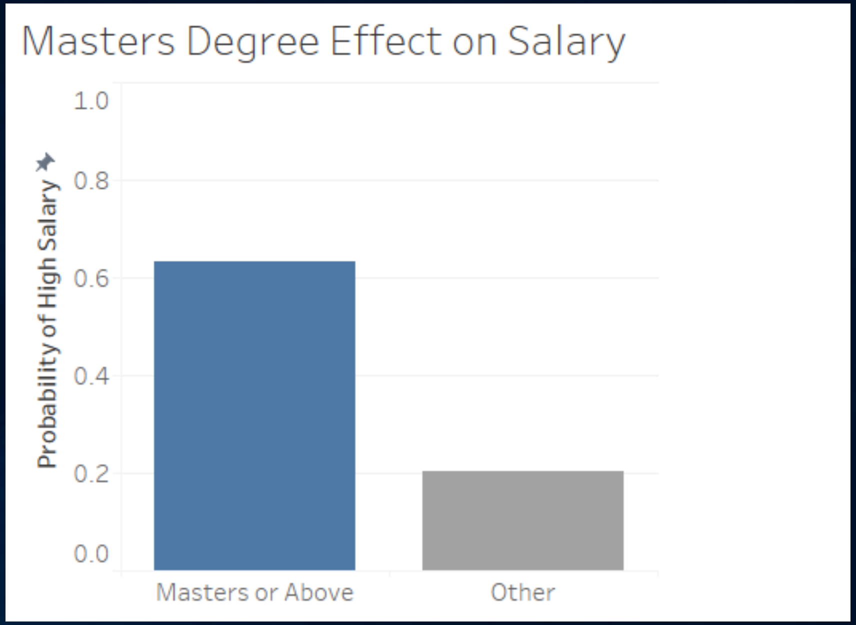


Figure 2

Best Indicators of High Salary (cont'd)

Figure 3

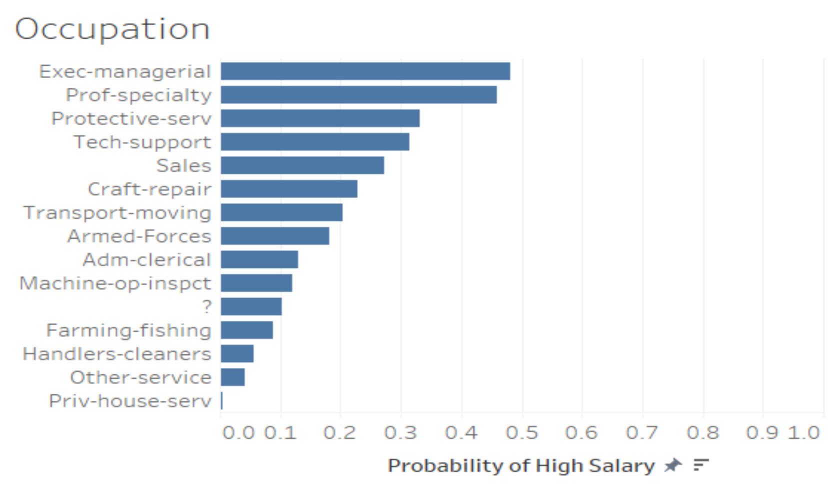
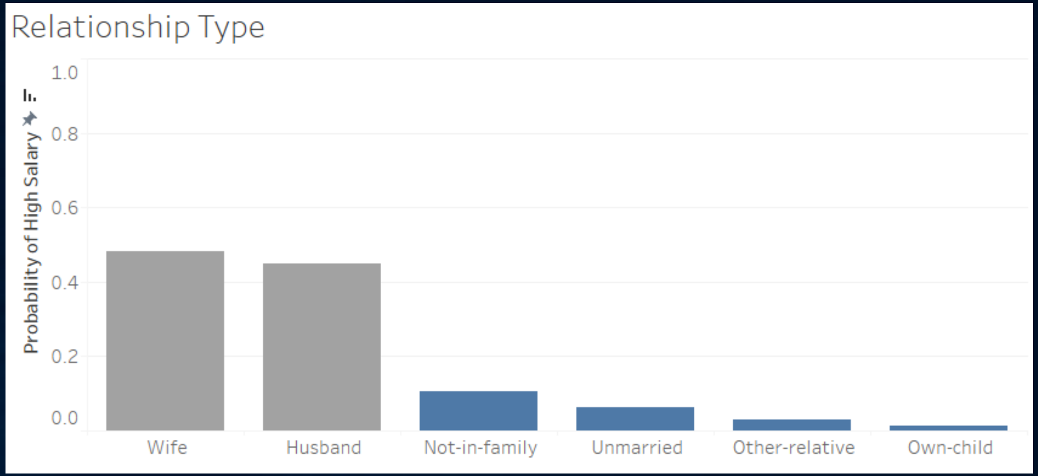


Figure 4



- **Occupation**
  - High salary widely varies among job types.

- **Relationship / Marriage**
  - Those who are a husband or wife have a 44% higher probability of making over $50k

Recommendations

1. Which specific types of people should be targeted?
   - An ideal customer would be:
     - Invests in the stock market (gains preferred)
     - Highly educated with at least a master's degree
     - Household role as a husband or wife
     - Occupation that is executive/managerial or specialty profession
2. Of those identified as ideal customers, spend more on individuals with the greatest probability of having a higher salary
3. Re-build the model by city or state, as outcomes could vary by region

Conclusion
Should the model be used?

PROS

- Best accuracy of all tested models
- Lowest chance of making a costly prediction out of all tested models
- Provides the ability to rank customers by probability of having a high salary

CONS

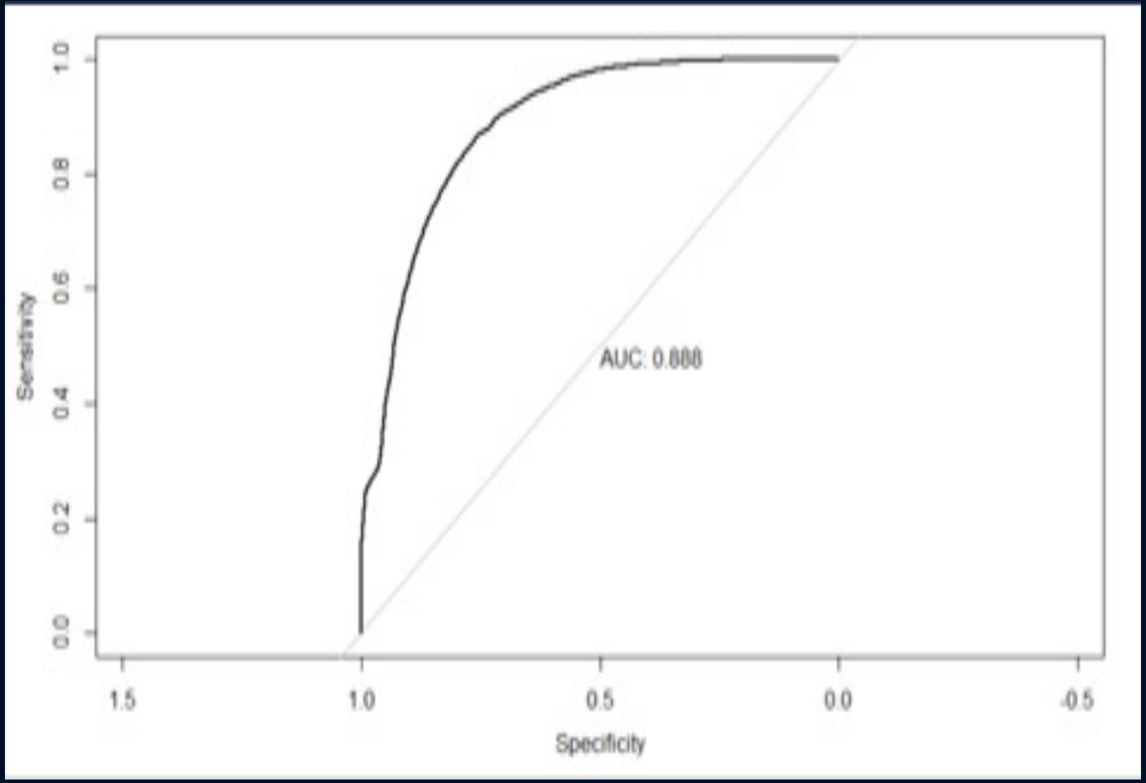
- Census data may not accurately represent bank's customer-base
- Chance of over-fitting
- Population across regions could have a different mix of characteristics than what was modeled

Thank you

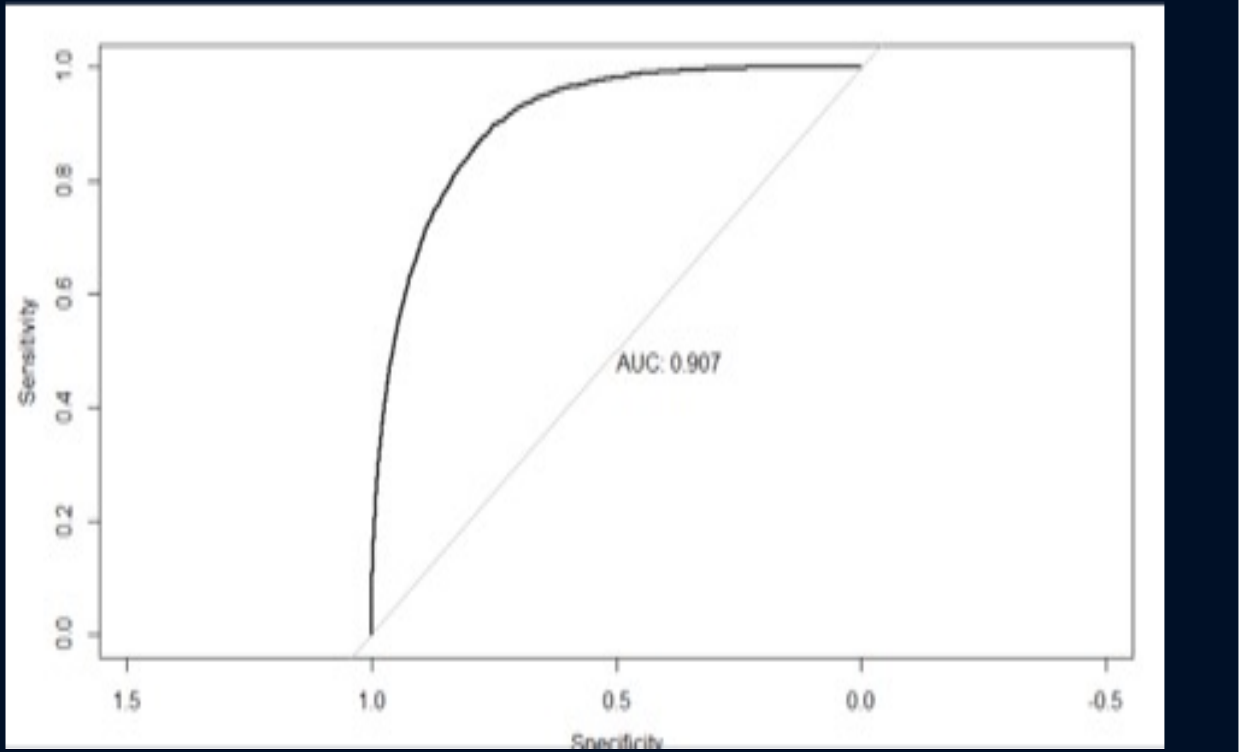
Appendix A - Model Comparisons

- Naïve Bayes model has an accuracy rate of 83% and approximately 53% of the correct predictions were for high income individuals.
- Logistic Regression model has an accuracy rate of 85% and around 60% of all the correct predictions were for high income group.
- Cart model has an accuracy rate of 85% and around 50% of all the correct predictions were for high income group.
- Random Forest model has an accuracy rate of 94% and around 86% of all the correct predictions were for high income group.

Appendix B - ROC Curves and AUC for all models

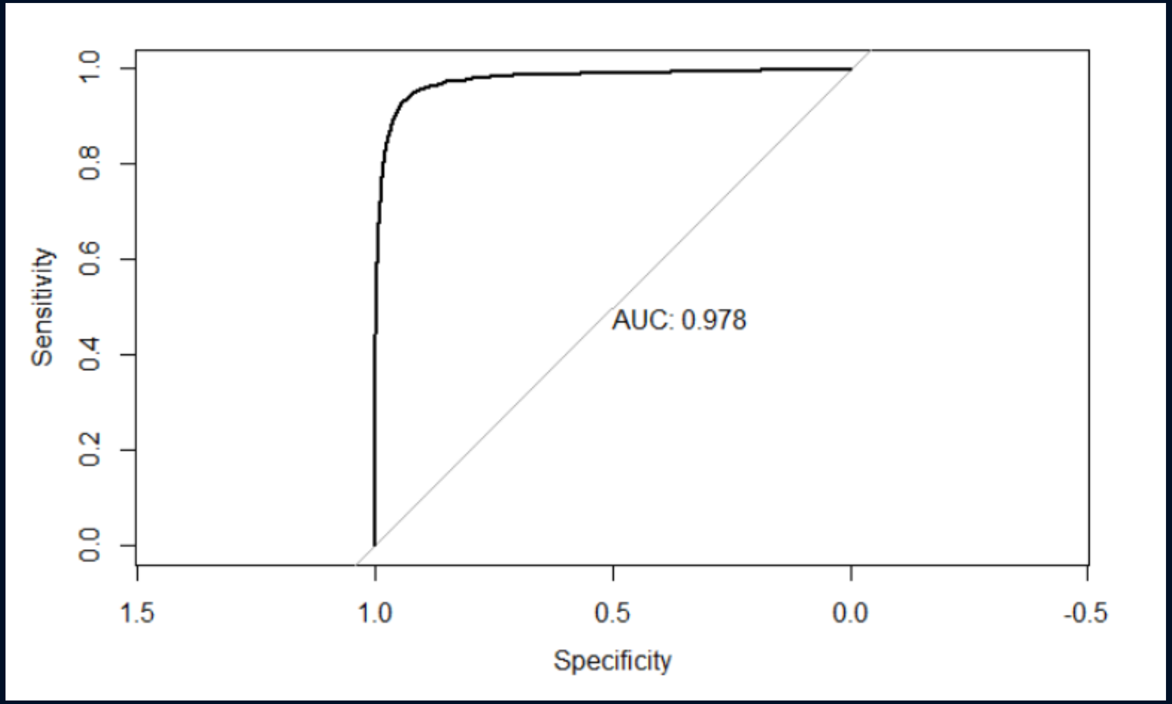


Naïve Bayes

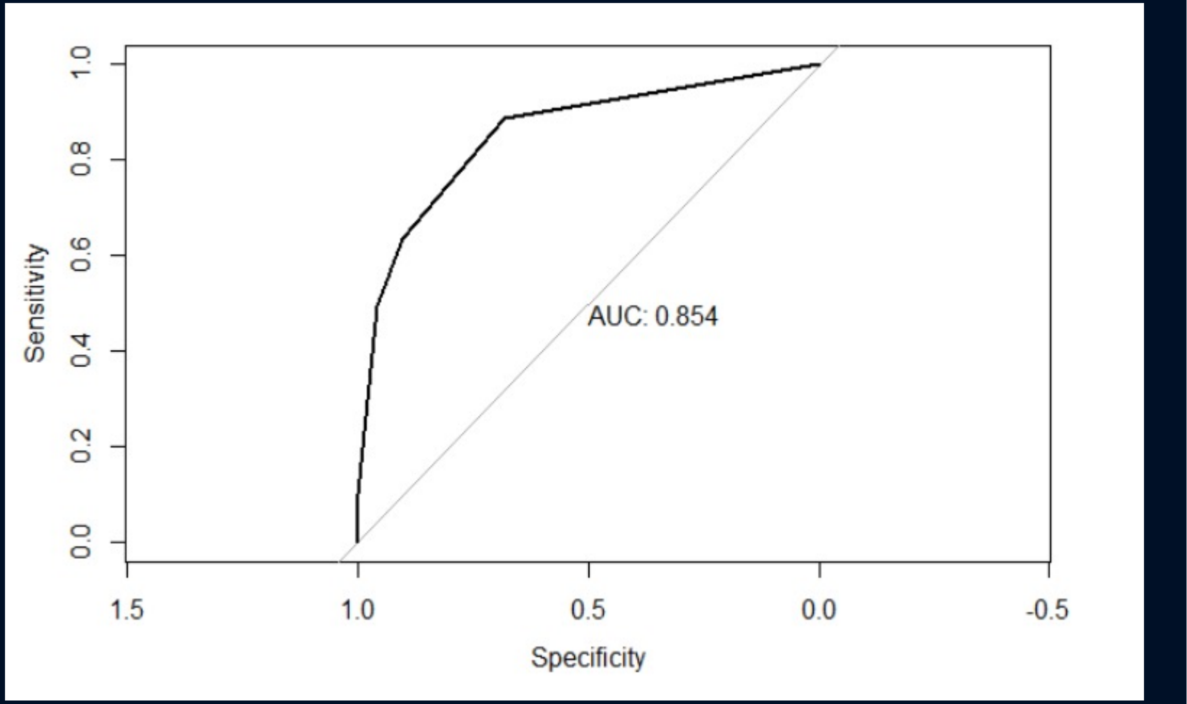


Logistic Regression

Appendix B (Cont'd)



Random Forest



Cart