Car Insurance Modeling

What is the problem?

- Our project aims to create a model which helps an insurance company decide what rate to charge new customers
- Our goal is to have our model predict whether a customer will file a claim
- A customer's rank can change depending on whether the model predicts they will file a claim or not
- Customers will be charged a rate based on their rank

The Data Set

- The data set contains 1000 rows and 19 columns of customer demographic information as well as vehicle type and driving behavior information

- Our target variable is whether or not someone filed a claim (identified in the “OUTCOME” column by a factor of 0 or 1)

- Our data set was imbalanced - 6,867 did not submit a claim, while 3,133 did

The Data Set

ID	Unique identifier; integer
AGE	Segmented
GENDER	male or female
RACE	majority or minority
DRIVING_EXPERIENCE	0-9 years, 10-19 years, 20-29 years, and 30+
EDUCATION	high school, university, or none
INCOME	middle class, upper class, poverty, or working class
CREDIT_SCORE	integer
VEHICLE_OWNER	0 or 1 (whether they own the vehicle)
VEHICLE_YEAR	after 2015, before 2015
MARRIED	0 or 1
CHILDREN	0 or 1
POSTAL_CODE	integer
ANNUAL_MILEAGE	integer
VEHICLE_TYPE	sedan or sports car
SPEEDING_VIOLATION	integer
DUIS	integer
PAST_ACCIDENTS	integer
OUTCOME	0 or 1 (whether they filed a claim)

Data Analysis

Correlations

We found that "ZIP_CODE" and "RACE" did not correlate

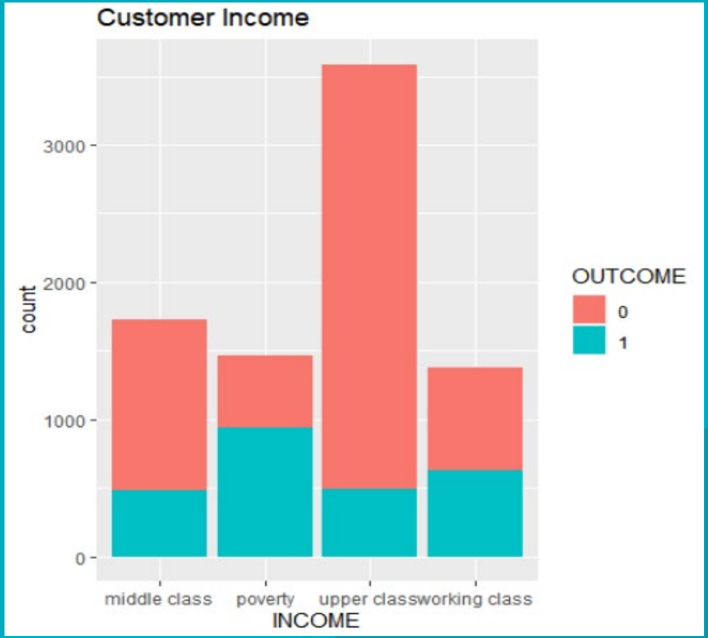
Filed

Most had between 0-9 years of driving experience and fell within the "poverty" category

Not Filed

Most had between 10-19 years of driving experience and were upper- class

Data Visualization



Higher income customers are less likley to submit a claim

Customers with more driving experience are less likely to submit a claim

Preprocessing

- We removed "GENDER" and "RACE" variables to minimize bias
- We removed the "ID" column and an additional 1,939 rows that contained missing information
- We added a column for rank so that customers with no accident were ranked "good", customers with 1 accident were ranked "okay", and customers with more than 1 accident were ranked "bad"
- After preprocessing there were 16 columns and 8,149 variables (5,613 did not submit a claim, while 2,536 did)

OUR MODELS

Decision Tree

Naive Bayes

Decision Tree

ACCURACY

83%, which was better than the no information rate of 68%

BALANCED ACCURACY

83%, the same as total accuracy

SENSITIVITY

83%, the model was good at classifying "no claims" correctly

PRECISION

83%, the model was good at classifying "claims" correctly

Fine Tuning

- For our project is was more important for us to accurately identify claims
- We tuned the model to place more weight on classifying claims

Fine Tuning

9 3 %

Accuracy increased for precision

6 8 %

Accuracy decreased for sensitivity

7 6 %

Total and balanced (80%) accuracy decreased

Confusion Matrix

		Actual	
Predicted			1
		2297	102
	1	1070	1419

Naive Bayes

ACCURACY

76%, which was better than the no information rate of 68%

BALANCED ACCURACY

74%, which is less than the total accuracy

SENSITIVITY

89%, the model was good at classifying "no claims" correctly

PRECISION

59%, the model was good at classifying "claims" correctly

Conclusion

1

The Decision Tree model out-performed the Naive Bayes model

2

Before tuning: driving experience, vehicle ownership, and age had the greatest impact

3

After tuning: driving experience, age, and past accidents had the greatest impact

Thank You!