DATA SCIENCE CAREER TRACK

HEALTH INSURANCE CROSS SELL PREDICTION

Damilola G. Efuwape (November, 2021)

Overview

Auto insurance, which is a contract between two parties, the insurance company and the insured customer, protects the insured against financial loss in the event of an accident or theft, as listed in the policy, in exchange for a premium being paid by the customer. In recent years, due to sporadic increase in vehicle owners, auto insurance has become very popular as the most profitable type of insurance next to health insurance. Most states in the United States require basic auto insurance for vehicles to be operated.

Many insurance companies provide different types of insurance and customers may get their different insurance coverages from one company or different ones for each insurance type. It is important for these companies to understand and be able to identify which of their customers would be willing to get a different kind of insurance with them and how they can get the customers to do so. One of the ways to do this could be offering each existing customer discounted rates if they will buy these other insurance coverages.

This study will take a look at the existing relationships among gender, age, sales channel and number of days the customers have been associated with the company to predict their potential responses.

Problem Statement

Which health insurance customers from the past year are likely to be interested in vehicle insurance?

Audience

The project is intended to help management and decision-makers at insurance companies identify potential auto insurance customers from their pool of already existing medical insurance customers.

Data

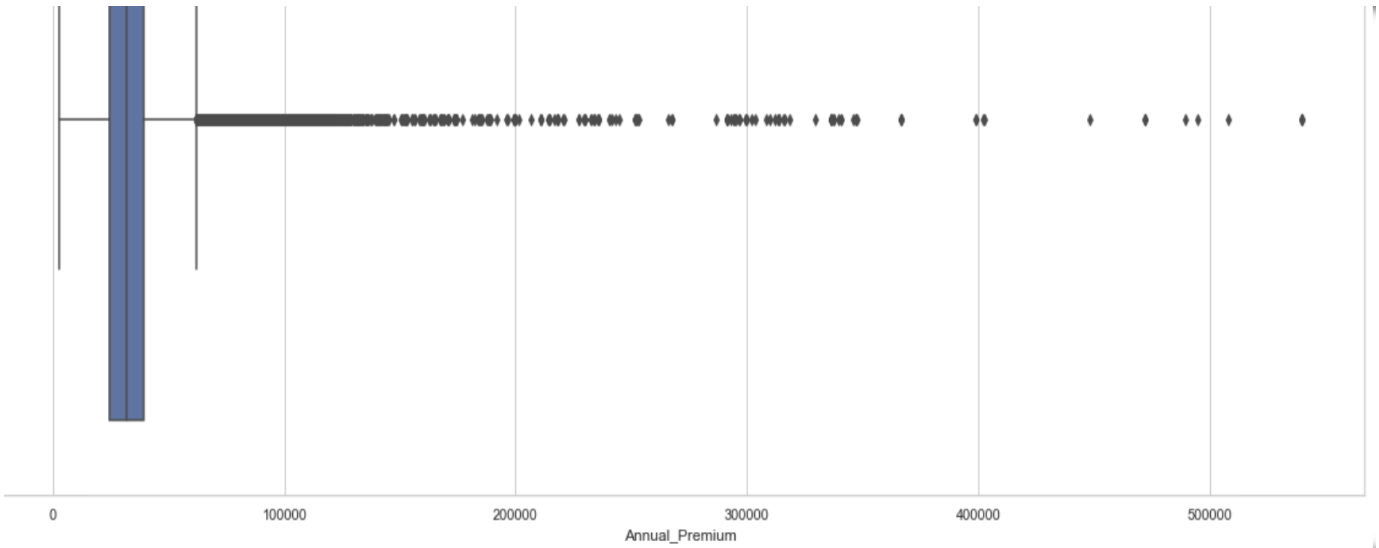
CSV-formatted health insurance test and train data with combined rows of five hundred and eight thousand, one hundred and forty-six and twenty-three columns from Kaggle will be used for this project. The dataset gives detailed information about the customers, as well as vehicular and policy information. The data usability is 10.0. Here is a link to the dataset.

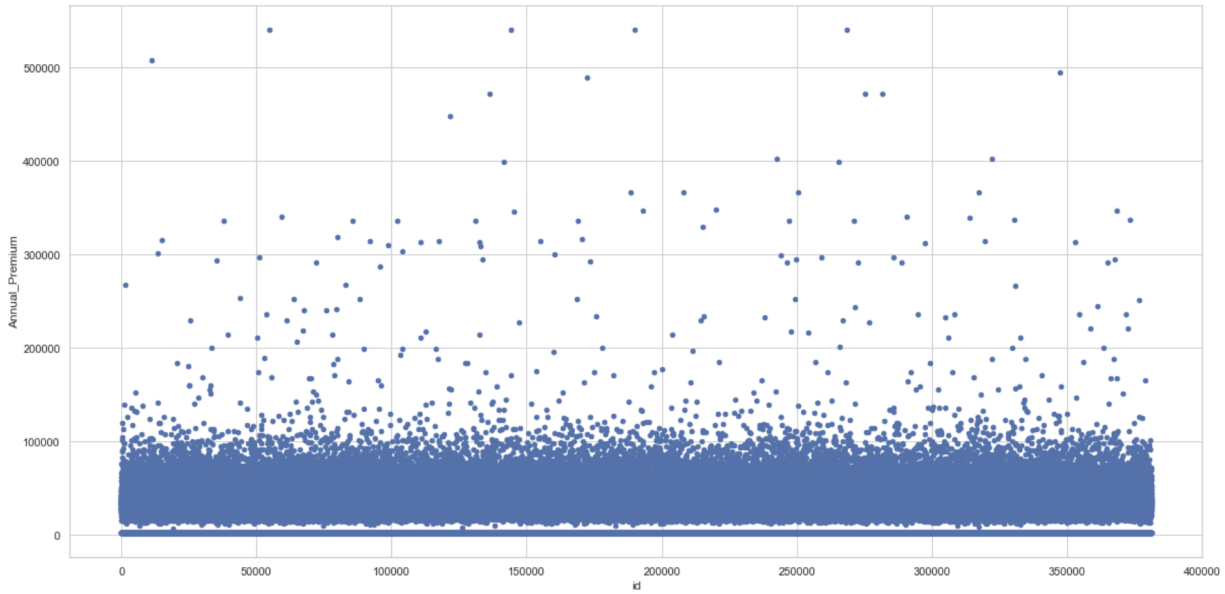
Data Wrangling

After importing the necessary packages and data, I created dummy variables for all categorical features contained in the data. I represented “Male” with “1” and “Female” with “0” for “Gender”. I also replaced “Yes” and “No” with “1” and “0” respectively for “Vehicle_Damage”. For “Vehicle_Age”, I put “0”, “1” and “2” in place of “< 1 Year”, “1-2 Year” and “> 2 Years” respectively.

Exploratory Data Analysis

I started this section by checking if there are any missing values and what percentages of each column are missing. If missing values existed in the data, I wanted to order them in increasing order and then present them in a single table. I found that there were no missing values anywhere in the data, so I went on to check for outliers after checking for duplicate values and found none. I discovered that the only outliers in the data were in the “Annual_Premium” column. This was confirmed through visualizations using a boxplot and a scatter plot as shown below:





These outliers did not have any effect on the analysis of the project because “Annual_Premium” was eventually dropped from the list of target features that were used for model building and deployment.

Next, I did some visualizations to see the relationships between features and check proportions of the different classes in the data. The following are some of the visualizations that were done and what they represent:

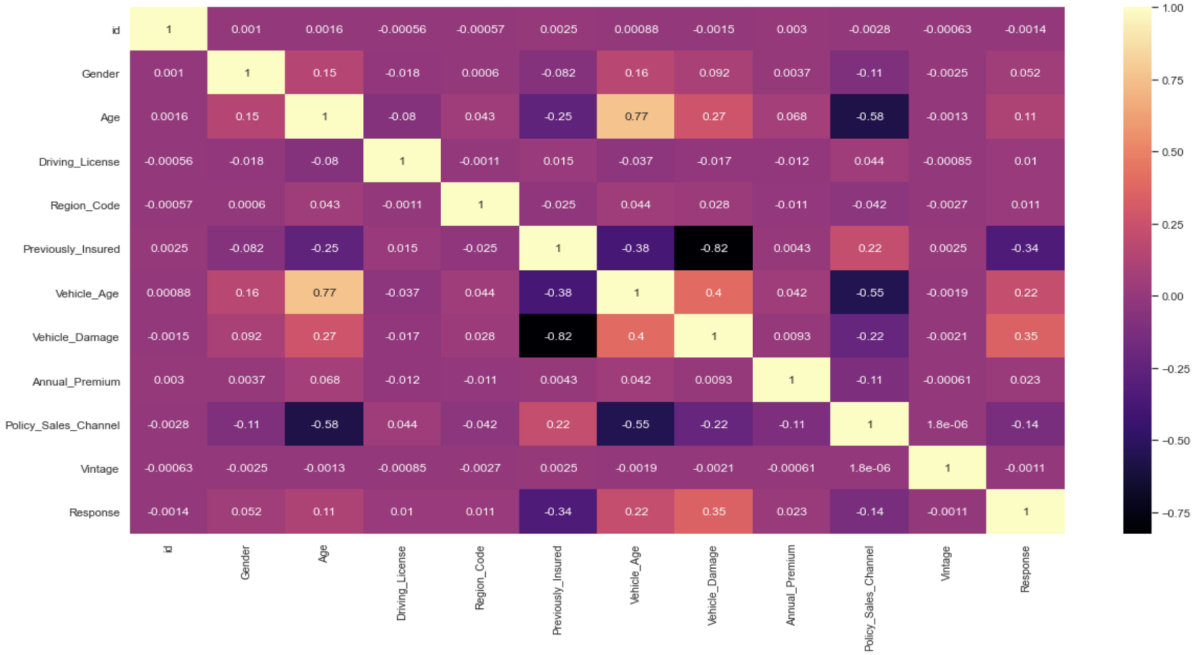


Fig. 1

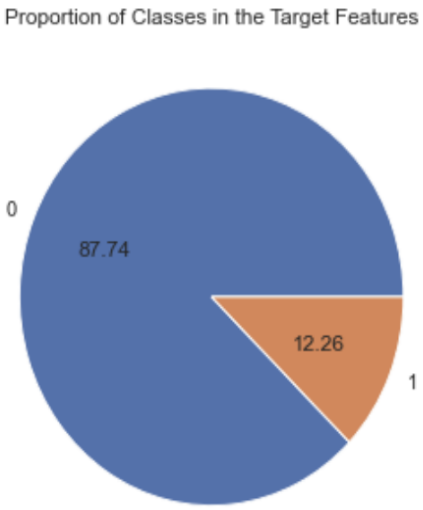


Fig. 2

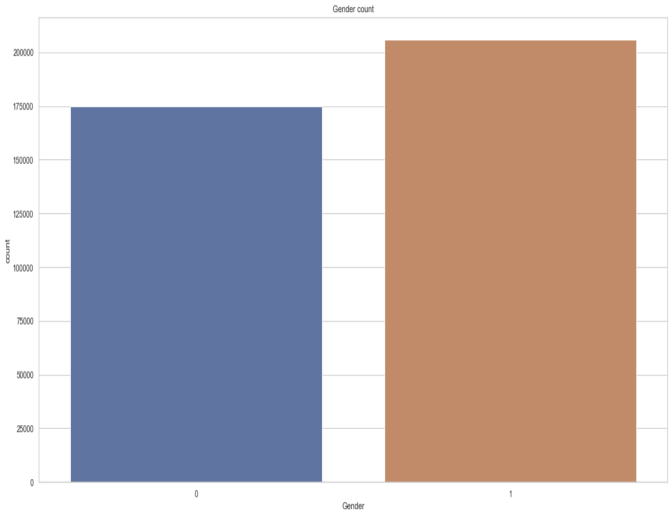


Fig. 3: Gender

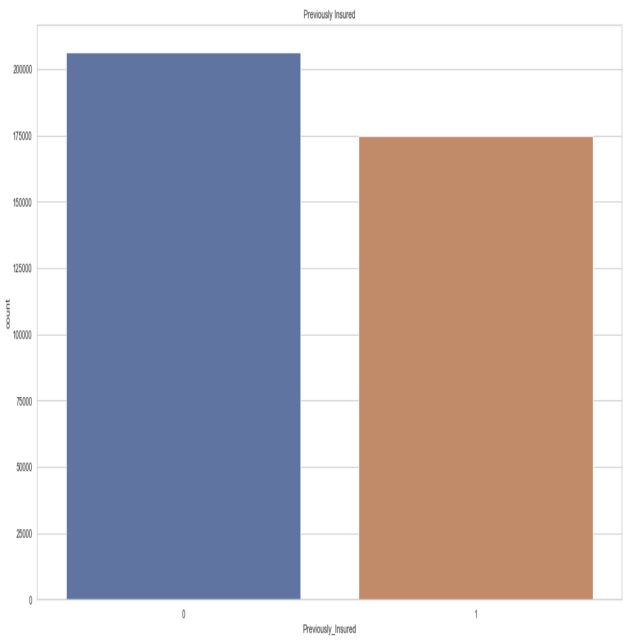


Fig. 4: Previously Insured

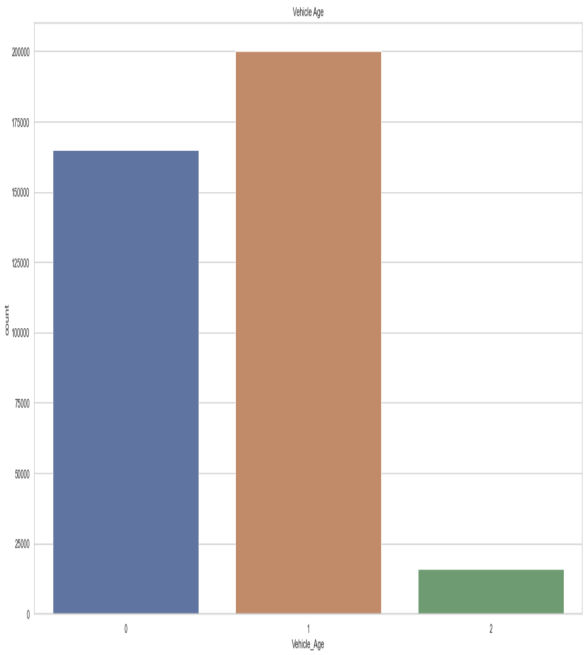


Fig. 5: Vehicle Age

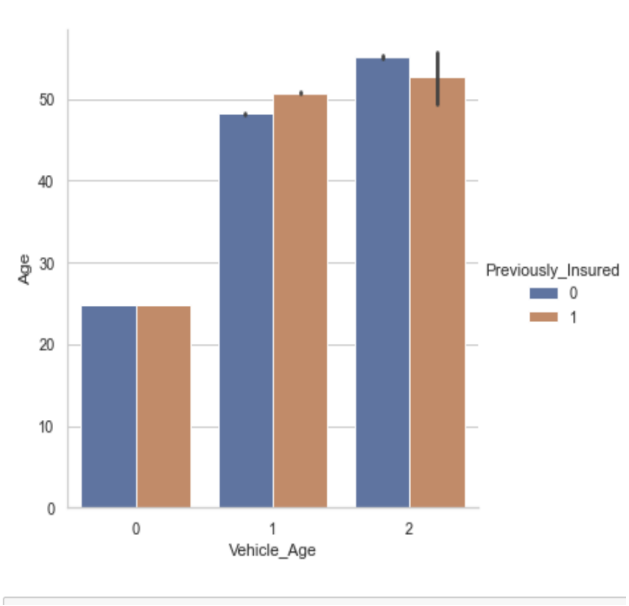


Fig. 6: Previously Insured Based On Vehicle Age

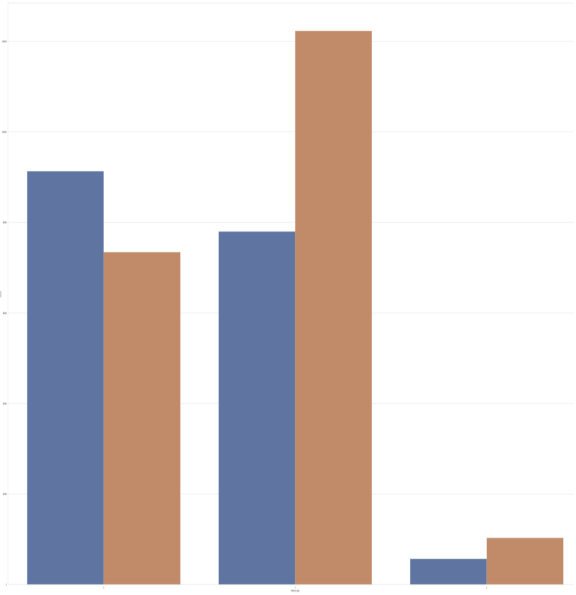


Fig. 7: Vehicle Age Based On Gender

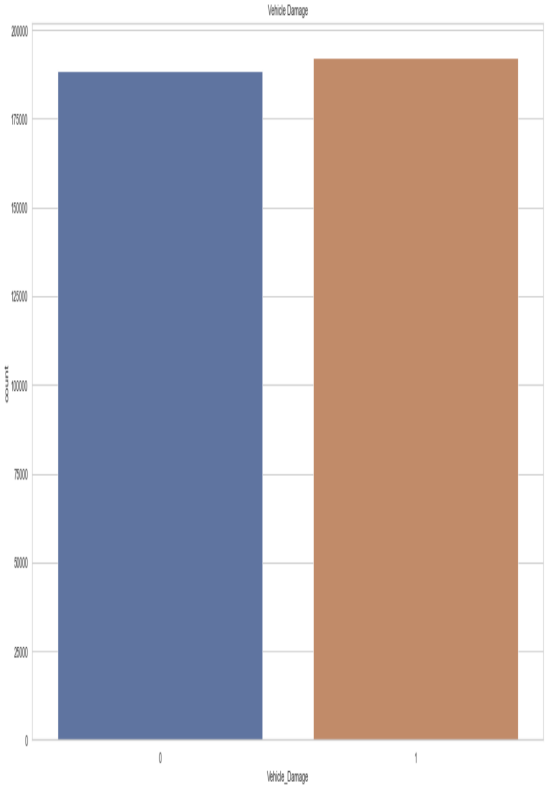


Fig. 8: Vehicle Damage

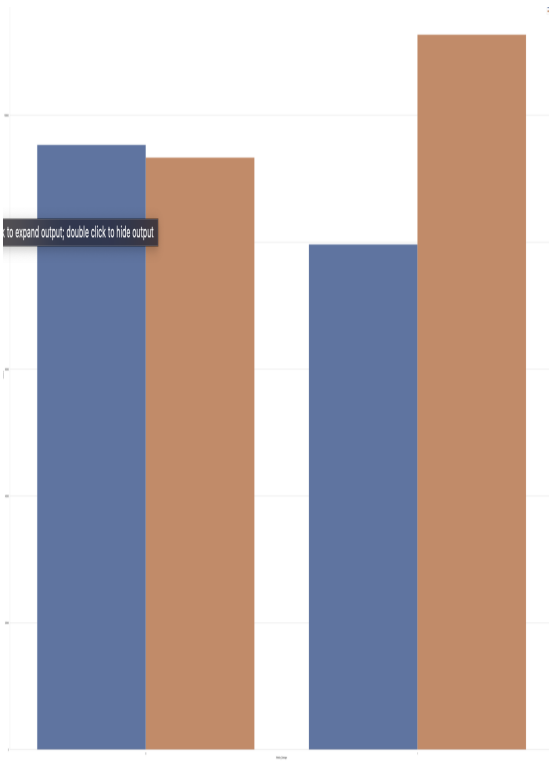


Fig. 9: Vehicle Damage Based On Gender

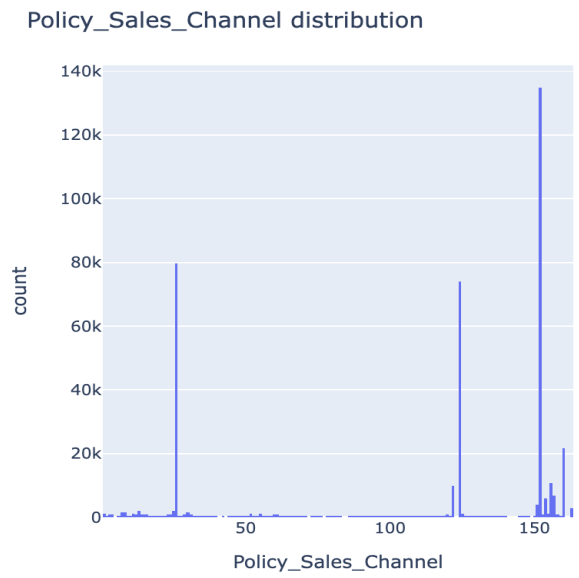


Fig. 10: Sales Channel

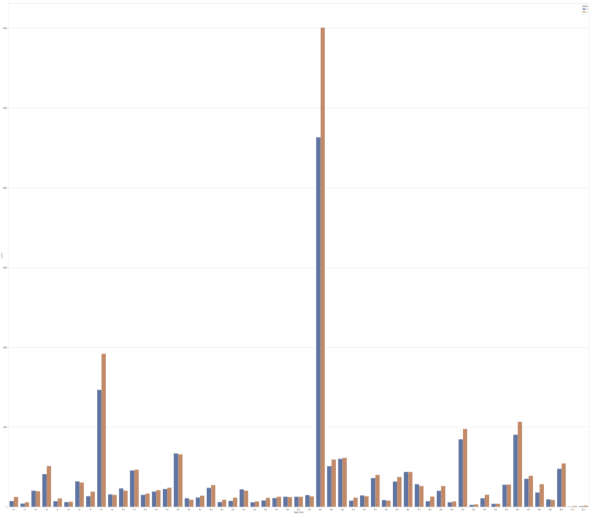


Fig. 11: Region

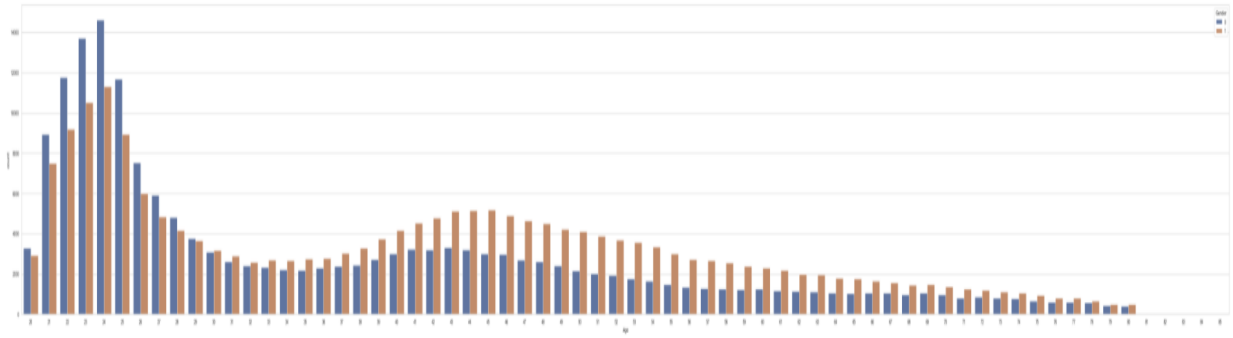


Fig. 12: Vehicle Owners Based On Age

Figure 1 shows the relationship among the features.

Figure 2 shows the proportion of classes in the target features. Just around 13% of the total customers gave a positive response as to if they are interested in getting auto insurance from the company.

Figure 3 shows that there are almost as many males(54%) as females(almost 46%) among our customers.

Figure 4 shows that there are slightly more uninsured(54%) people among the customers than the insured ones(~46%). These could be considered potential customers.

Figure 5 shows that most of the customers use vehicles that are aged between one and two years, followed by less than a year old vehicles. Very few of them use vehicles that are over two years old.

Figure 6 shows that the proportions of insured customers to their uninsured counterparts based on vehicle age are almost the same.

Figure 7 shows that most of the customers who own vehicles that are less than a year old are females and males are top in the other two categories, probably because there are slightly more males than females among the customers.

Figure 8 shows that there are almost as many customers who have had their vehicles damaged as those who have not.

Figure 9 separates those who have damaged their vehicles and those who have not into male and female categories. It shows that more males have damaged their vehicles than females. Are females more careful than males?

Figure 10 shows that sales channels 152, 26 and 124 brought in the highest, second highest and third highest numbers of customers respectively.

Figure 11 shows that for most region code areas, the numbers of male customers are pretty close to those of females, with regions 28 and 8 having the highest and second highest numbers of customers respectively.

Figure 12 shows that more females tend to own vehicles at a younger age than males. But, this changes as the ages increase, that is, more older males own vehicles than older females.

From the above, it can be inferred that most customers are aged between 20 and 30. Almost everyone has a driver's licence. Areas with region codes 28 and 8 have the highest and second highest represented customers respectively. It could mean these are high income areas and people there can afford to pay for insurance. Most of the customers got their insurance with the company through sales channels somewhere between 20-30, 120-130 and 150-160 (26, 124 and 152). Could these channels be more appealing to the customers, especially young people, since they make up most of the customers?

I ended this section by performing some data profiling to get comprehensive general information about the whole data.

Modeling

For effective prediction, eight models were built and deployed. These models cut across linear regression, probability, tree-based and boosting techniques. The models are Linear Regression, Naive Bayes (Gaussian), Naive Bayes (Bernoulli), XGBoost, LightGBM, Gradient Boosting, Random Forest and Decision Tree.

The evaluation metric for these models is the Receiver Operating Characteristic, Area Under Curve (ROC_AUC) score. The higher the score the better the accuracy of the model. However, due to the high imbalance in the classes of the target features, great importance was given to the models' F1 scores on the test data.

After deployment to the original data, the Naive Bayes (Bernoulli) had the highest F1 score (0.368) on the test data with an average F1 score of 0.365 during cross validation, which indicates no signs of overfitting.

Decision Tree and Random Forest are the second and third best performing models with F1 scores of 0.302 and 0.255; and mean F1 scores of 0.300 and 0.254 respectively.

All other models performed really bad, with very negligible scores across train and test data.

Tree based and boosting based algorithms suffered heavily due to heavy imbalance in target classes across train and test where the proportion of class 0 to 1 is about 87% to 13%.

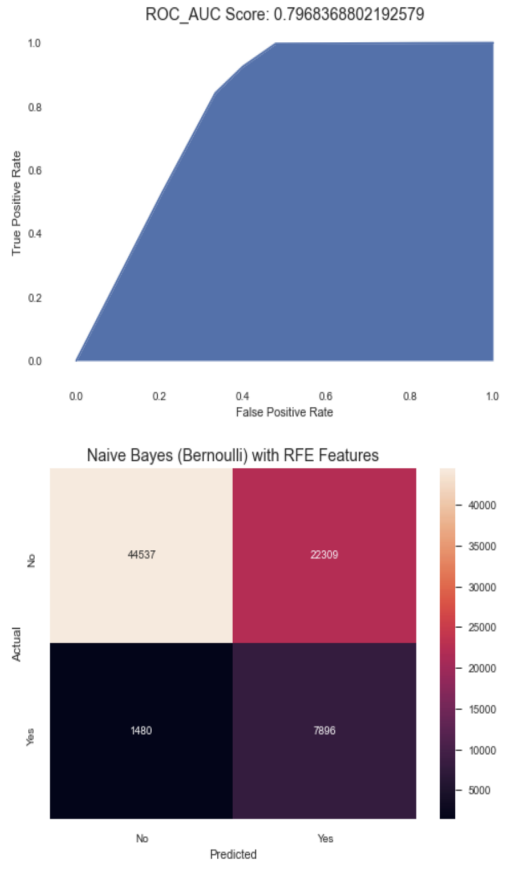
Resampling was then performed on the training data, especially the minority class, to deal with the imbalance of the data and improve the scores.
After resampling and redeployment of the three best performing models from the previous deployment, the Naive Bayes (Bernoulli) model, again, gave the highest F1 scores of 0.667 and 0.420 on training and testing data respectively, with a mean F1 score of 0.667, although it produced very low precision for class 1 on test data, which is, 0.28 compared to recall of 0.83.

Random Forest and Decision Tree, which both had slightly lesser differences in recall and precision scores in test samples, were second and third best performing models in terms of F1 respectively.

Tree-based algorithms are prone to overfitting and the deviation in cross validation scores was close to 1.7% for Decision Tree and 1.3% for Random Forest compared to just 0.3% by Naive Bayes (Bernoulli), which generalized better and provided the most consistent results.

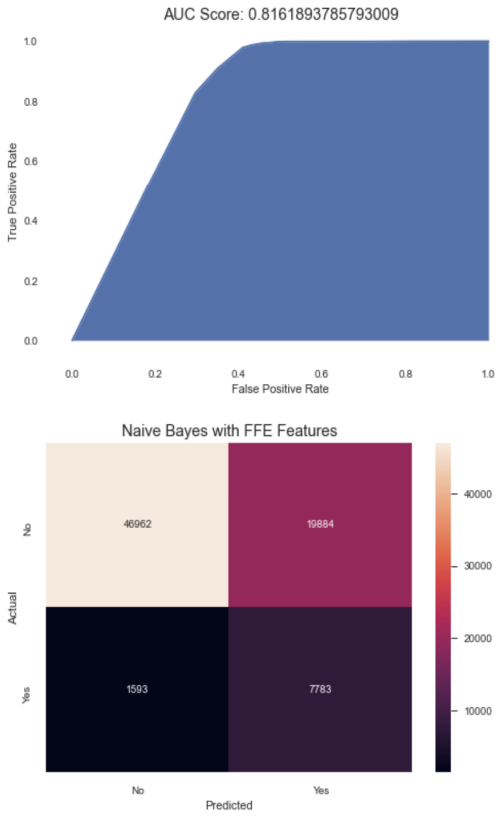
Since, Naive Bayes (Bernoulli) gave the highest F1 score, I chose the Naive Bayes (Bernoulli) for further tuning.

To determine the best set of features for Naive Bayes (Bernoulli) during feature selection, I used Recursive Feature Elimination and Forward Feature Elimination.
The Naive Bayes (Bernoulli) model gave higher F1 scores of 0.667 and 0.420 on train and test samples respectively, using Forward Feature Elimination. The model, however, had a higher deviation in scores for Forward Feature Elimination (0.30%) than for Recursive Feature Elimination (0.28%) as shown in the graphs below:



Using Recursive Feature Elimination, Naive Bayes (Bernoulli) has an ROC_AUC score of 79.68.

Fig. 13: Naive Bayes' ROC Curve with RFE



Forward Feature Elimination yielded an ROC_AUC score of 81.62

Fig. 14: Naive Bayes' ROC Curve with FFE

	Model	Train F1	Test F1	CV Mean F1	CV std in scores
0	Recursive Feature Elimination	0.652751	0.398979	0.652756	0.0028211
1	Forward Feature Elimination	0.667006	0.420214	0.667011	0.00303703

It can be inferred from the diagrams above that the model yielded a higher ROC_AUC score of 81.62 for Forward Feature Elimination than for Recursive Feature Elimination (79.68).

Hence I picked the Naive Bayes (Bernoulli) model with a set of features determined by Forward Feature Elimination as the best model to be used to identify already existing medical insurance customers who will be willing to purchase an auto insurance through the company.

Finally, the model was saved and deployed to the test data to show customers' likely responses and the some of the responses are shown below:

	id	Response
0	381110	0
1	381111	0
2	381112	0
3	381113	0
4	381114	0
...	...	...
127032	508142	0
127033	508143	0
127034	508144	0
127035	508145	0
127036	508146	0

127037 rows × 2 columns

Business Interpretation of the Model

Inferences from the whole analysis show that more males gave a positive response than females. Customers aged between mid-thirties and mif-fifties also gave a positive response. Customers who have driver's licences are most likely to give a positive response (customers who do not have driver's licences are not many anyway). Previously insured customers all gave a negative response, so those without vehicle insurance are our best bet and are a better target. Customers with vehicles aged between one and two years gave more positive responses than those in the other categories. Majority of customers who have not damaged their vehicles said they are not interested in our vehicle insurance, which might make us shift our attention to those who have, at one time or the other, damaged their vehicles. Customers who got their insurance through channels 152, 26 and 124, especially those from regions 28 and 8, are most interested in our auto insurance compared to those who got their insurance through other channels. How long a customer has been patronizing the company (Vintage) and annual premium he or she pays do not affect their responses.

Overall, only about 13% of over 380,000 health insurance customers in the original dataset gave a positive response as to whether they would like to purchase an auto insurance through the company. SUCH IMBALANCE!

After in-depth analysis of all features, I have created a Naive Bayes (Bernoulli) model to predict which health insurance customers would be interested in purchasing vehicle insurance from the company. The following are key insights about the model:

The most useful features for identifying potential vehicle insurance customers were chosen to create the models in this project. These features can be the major target points for the company when trying to identify potential auto insurance customers. The used features include:

1. Vehicle Damage.
2. Age.
3. Previously Insured.
4. Gender
5. Region Code, preferably 28, 152 or 8.
6. Policy Sales channel, preferably 26, 124 or 152.
7. Vehicle Age.
8. Customer Id

Based on the chosen target features, the final model had an F1 score of 0.67 on the training data and an F1 score of 0.42 on test data, with an ROC_AUC score of 81.62. The model was very reliable and produced consistent results, based on a deviation in f1 scores across several test samples calculated to only about 0.3%.

During testing, the model gave an AUC score of 81.62, which is a good score.

The project was not without some challenges though as the model experienced some limitations, amongst which is a huge difference in the test and train scores of the model, which indicates underfitting and was a problem across almost all models that were built due to the high imbalance in instances of health insurance customers who have auto insurance and those who did not.

After resampling the data for the classes of customers in the training sample, the models performed better than they did on the original data with results for the test set significantly increasing.

However, while the scores during the test phase had improved significantly, there was still a huge difference in the train and test results.

Recommendation

So, based on the findings of this project, my recommendations are that male customers who have driver's licenses, who are between mid-thirties and mid-fifties, who have had their vehicles damaged at some point, who do not currently have their vehicles insured, whose vehicles are aged between one and two years, and got their insurance through channels 152, 26 and 124 especially those from regions 28 and 8, are our likely potential auto insurance customers.

Suggestion For Improvement

More data of health and auto insurance customers should be added to the data to avoid high data imbalance. This will greatly improve the model and help get better results during testing. Also, the level of contribution of customer age to vehicle damage could be checked. That is, what age groups have damaged the most and least vehicles. Some other sophisticated or high-level classification models should also be used on the data to see if they will make better predictions.