Health Insurance Cross Sell Prediction

An Analysis Using Machine Learning Techniques

Prepared By:

Umair Chaanda

Introduction

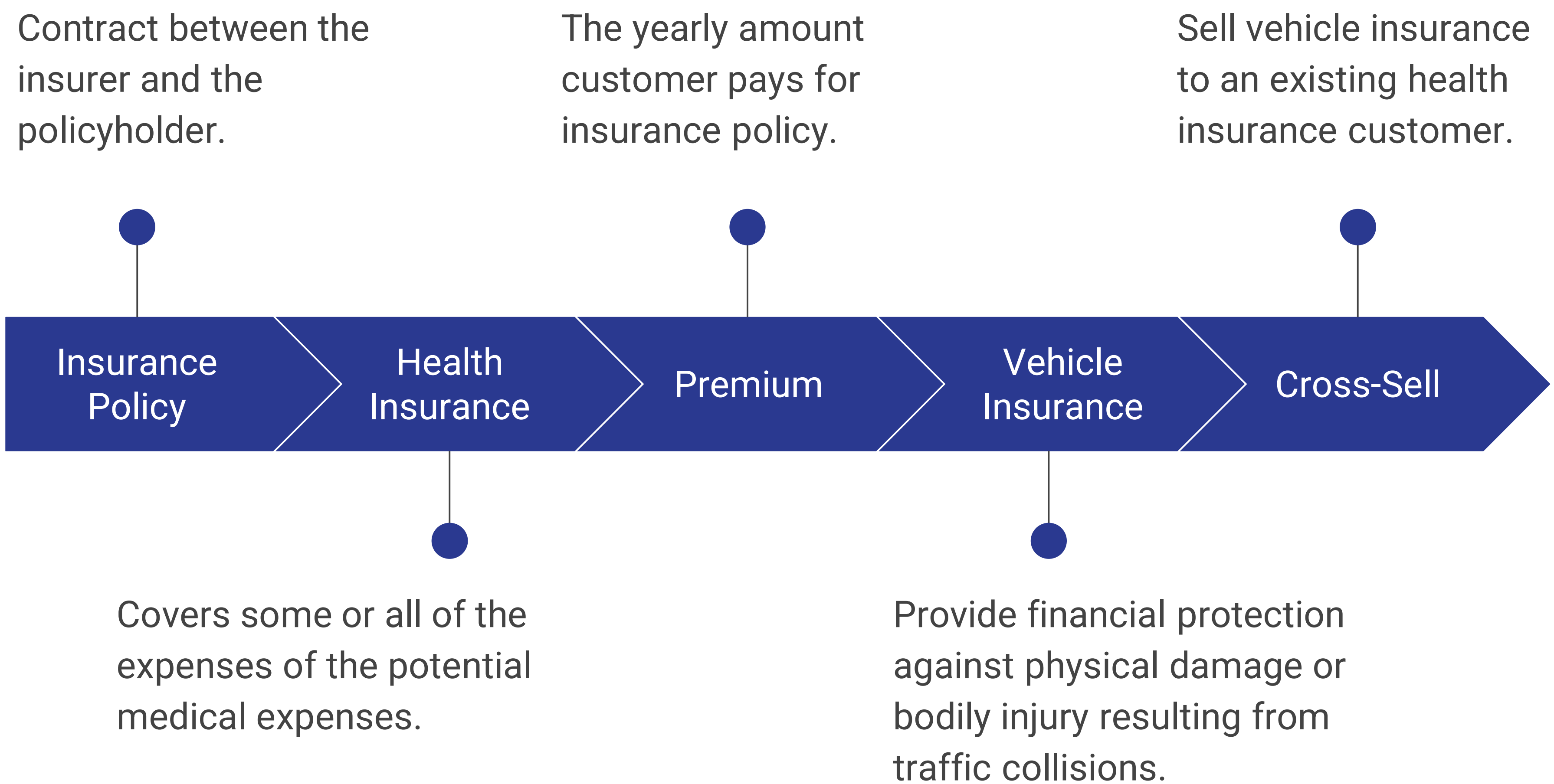
Cross-Sell

- Cross-selling is a frequently used marketing strategy.

- The chief purpose of cross-selling is to generate a positive revenue flow (from existing customers) by selling a variety of product lines.

Our Main Goal

- Predict whether the Health Insurance policyholders (customers) from past year will also be interested in Vehicle Insurance.
- Pure binary (1: Interested, 0: Not Interested) classification task.
- The predictive model is extremely helpful for the company to plan its communication strategy and optimise its business model and revenue.


Contract between the insurer and the policyholder.
The yearly amount customer pays for insurance policy.
Sell vehicle insurance to an existing health insurance customer.
Insurance Policy
Health Insurance
Premium
Vehicle Insurance
Cross-Sell
Covers some or all of the expenses of the potential medical expenses.
Provide financial protection against physical damage or bodily injury resulting from traffic collisions.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis

Cleaning & Preprocessing

- Data Description
- General Info - Train Data
- General Info - Test Data
- Check and handle Missing and Duplicate Data
- Convert Variable Types
- Data Transformation
- General Statistics of Features

Data Visualization

- Distribution of Categorical Features.
- Cross Tabulated View using Bar Plots
- Correlation Analysis
- Proportion of Target Variable
- Pairplot
- Catplots

Dataset

Publicly-available and acquired from Kaggle

Target Variable: Response

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy_Sales_Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

General Information About Data

Train

```
Int64Index: 381109 entries, 1 to 381109
Data columns (total 11 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Gender                381109 non-null  object
 1   Age                   381109 non-null  int64
 2   Driving_License       381109 non-null  int64
 3   Region_Code           381109 non-null  float64
 4   Previously_Insured    381109 non-null  int64
 5   Vehicle_Age           381109 non-null  object
 6   Vehicle_Damage        381109 non-null  object
 7   Annual_Premium        381109 non-null  float64
 8   Policy_Sales_Channel  381109 non-null  float64
 9   Vintage               381109 non-null  int64
 10  Response              381109 non-null  int64
dtypes: float64(3), int64(5), object(3)
```

```
Rows, Columns
(381109, 12)
```

Test

```
Int64Index: 127037 entries, 381110 to 508146
Data columns (total 10 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Gender                127037 non-null  object
 1   Age                   127037 non-null  int64
 2   Driving_License       127037 non-null  int64
 3   Region_Code           127037 non-null  float64
 4   Previously_Insured    127037 non-null  int64
 5   Vehicle_Age           127037 non-null  object
 6   Vehicle_Damage        127037 non-null  object
 7   Annual_Premium        127037 non-null  float64
 8   Policy_Sales_Channel  127037 non-null  float64
 9   Vintage               127037 non-null  int64
dtypes: float64(3), int64(4), object(3)
```

```
Rows, Columns
(127037, 11)
```

Data Transformation

FLOAT

INT

Region Code

Annual Premium

Policy Sales Channel

STR

Male	1
Female	0

Yes	1
No	0

INT

> 2 Years	3
1-2 Years	2
< 1 Years	1

Basic Statistics of Features

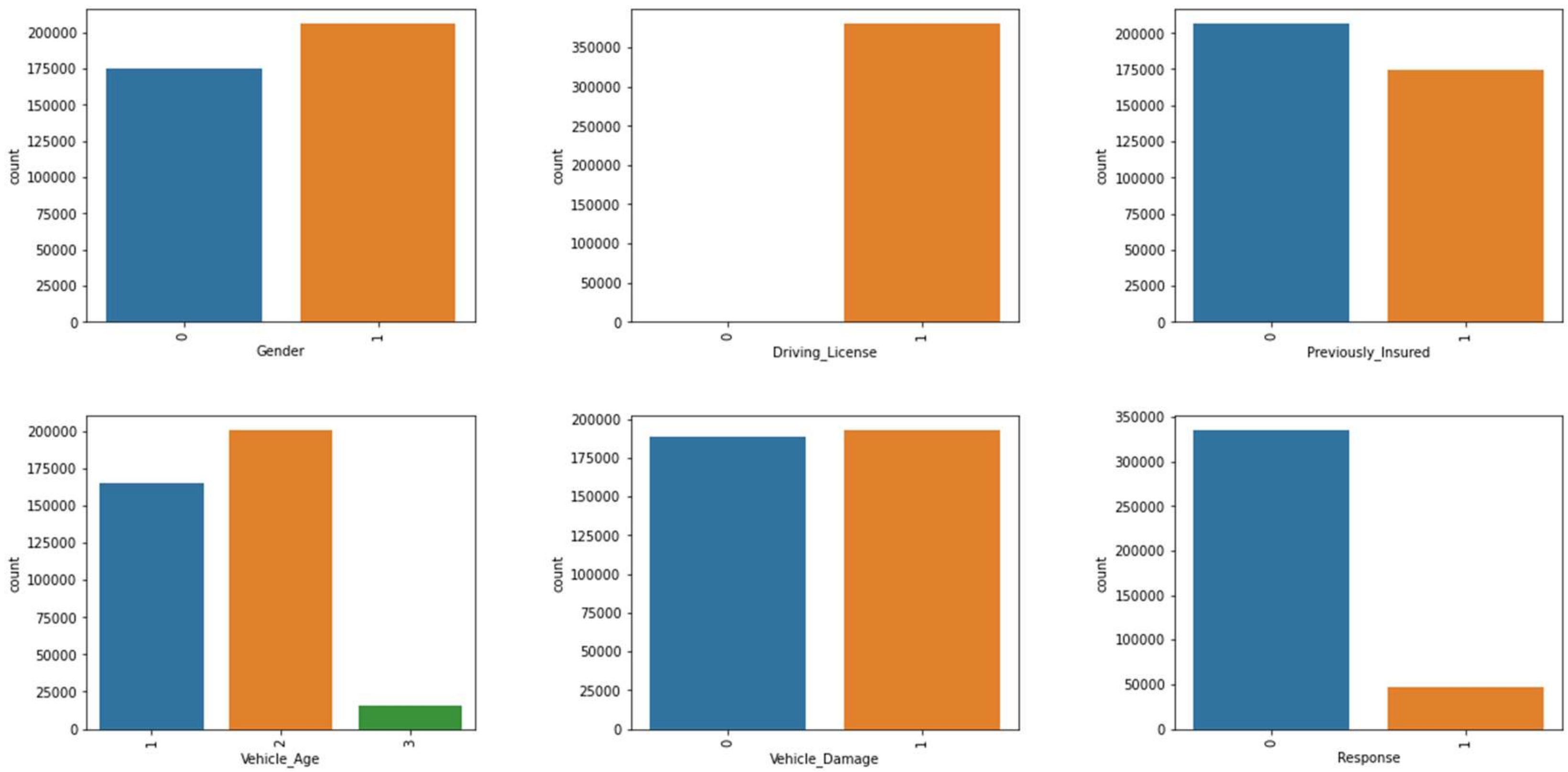
Train

	count	mean	std	min	25%	50%	75%	max
Gender	381109.0	0.540761	0.498336	0.0	0.0	1.0	1.0	1.0
Age	381109.0	38.822584	15.511611	20.0	25.0	36.0	49.0	85.0
Driving_License	381109.0	0.997869	0.046110	0.0	1.0	1.0	1.0	1.0
Region_Code	381109.0	26.388807	13.229888	0.0	15.0	28.0	35.0	52.0
Previously_Insured	381109.0	0.458210	0.498251	0.0	0.0	0.0	1.0	1.0
Vehicle_Age	381109.0	1.609616	0.567439	1.0	1.0	2.0	2.0	3.0
Vehicle_Damage	381109.0	0.504877	0.499977	0.0	0.0	1.0	1.0	1.0
Annual_Premium	381109.0	30564.389581	17213.155057	2630.0	24405.0	31669.0	39400.0	540165.0
Policy_Sales_Channel	381109.0	112.034295	54.203995	1.0	29.0	133.0	152.0	163.0
Vintage	381109.0	154.347397	83.671304	10.0	82.0	154.0	227.0	299.0
Response	381109.0	0.122563	0.327936	0.0	0.0	0.0	0.0	1.0

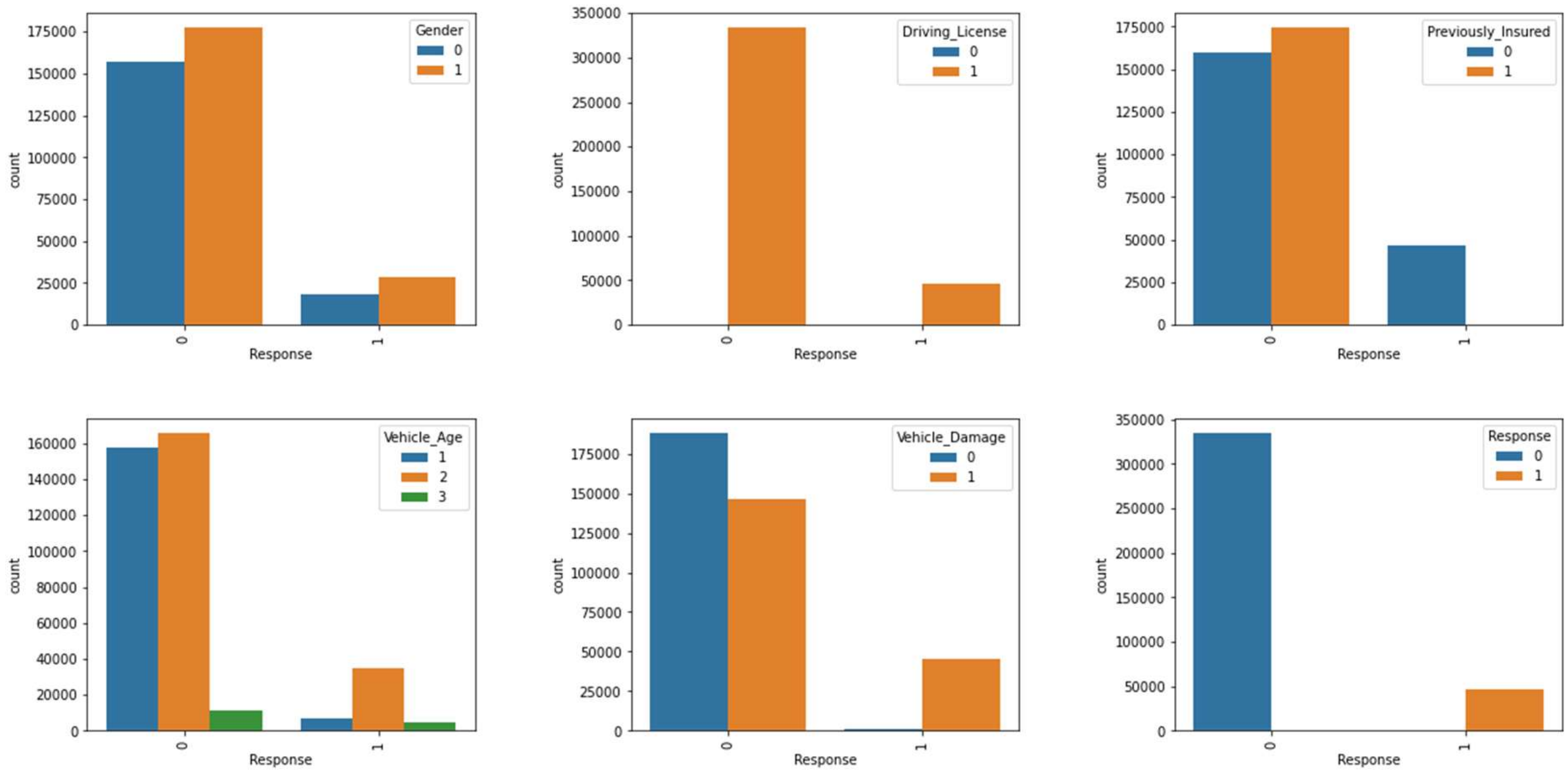
Test

	count	mean	std	min	25%	50%	75%	max
Gender	127037.0	0.537135	0.498621	0.0	0.0	1.0	1.0	1.0
Age	127037.0	38.765903	15.465814	20.0	25.0	36.0	49.0	85.0
Driving_License	127037.0	0.998134	0.043152	0.0	1.0	1.0	1.0	1.0
Region_Code	127037.0	26.459866	13.209916	0.0	15.0	28.0	35.0	52.0
Previously_Insured	127037.0	0.460039	0.498403	0.0	0.0	0.0	1.0	1.0
Vehicle_Age	127037.0	1.608775	0.567371	1.0	1.0	2.0	2.0	3.0
Vehicle_Damage	127037.0	0.502491	0.499996	0.0	0.0	1.0	1.0	1.0
Annual_Premium	127037.0	30524.643576	16945.297103	2630.0	24325.0	31642.0	39408.0	472042.0
Policy_Sales_Channel	127037.0	111.800468	54.371765	1.0	26.0	135.0	152.0	163.0
Vintage	127037.0	154.318301	83.661588	10.0	82.0	154.0	227.0	299.0

Distribution of Categorical Features



Cross Tabulated View



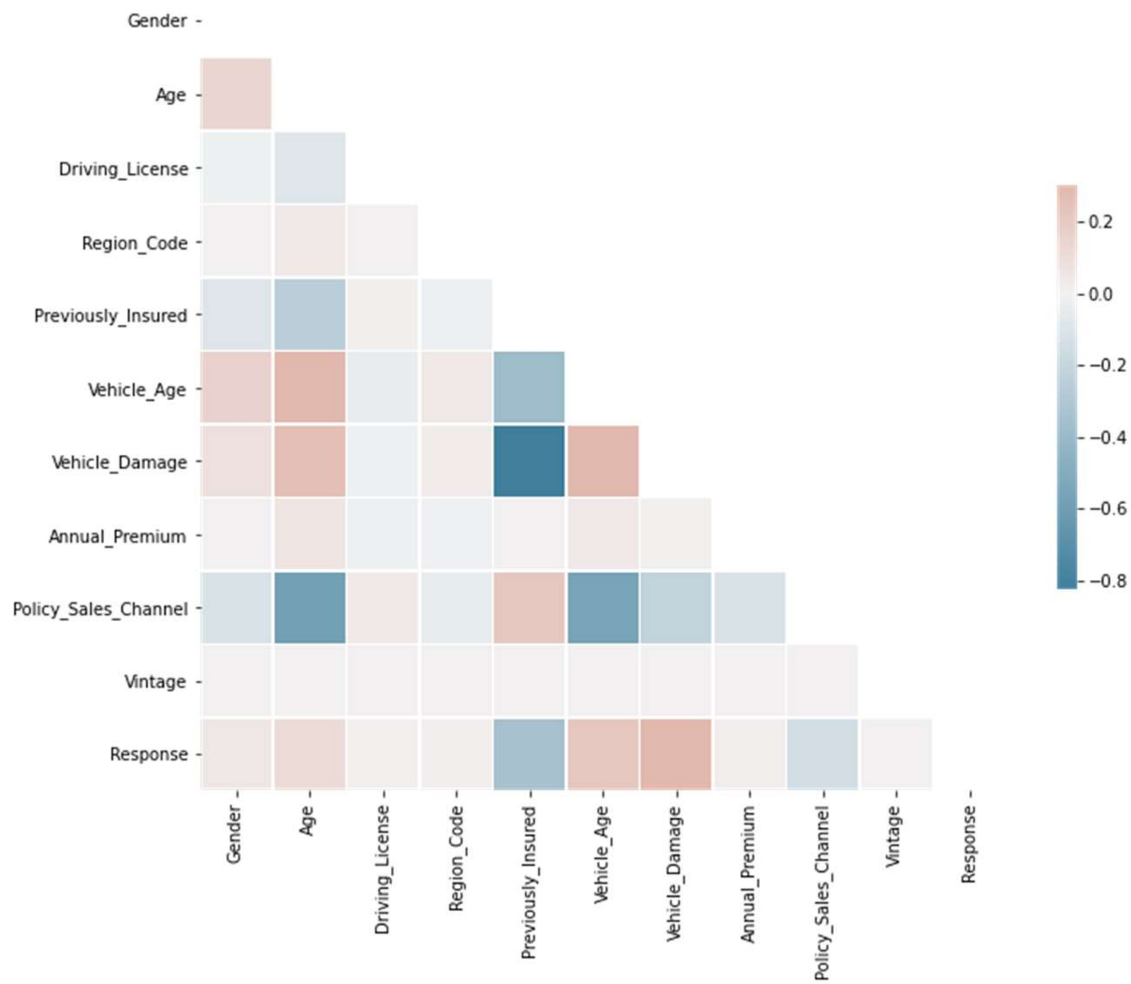
Correlation Analysis

Positively Correlated Features:

- Vehicle_Age and Age
- Vehicle_Damage and Age
- Vehicle_Damage and Vehicle_Age
- Vehicle_Damage and Response

Negatively Correlated Features:

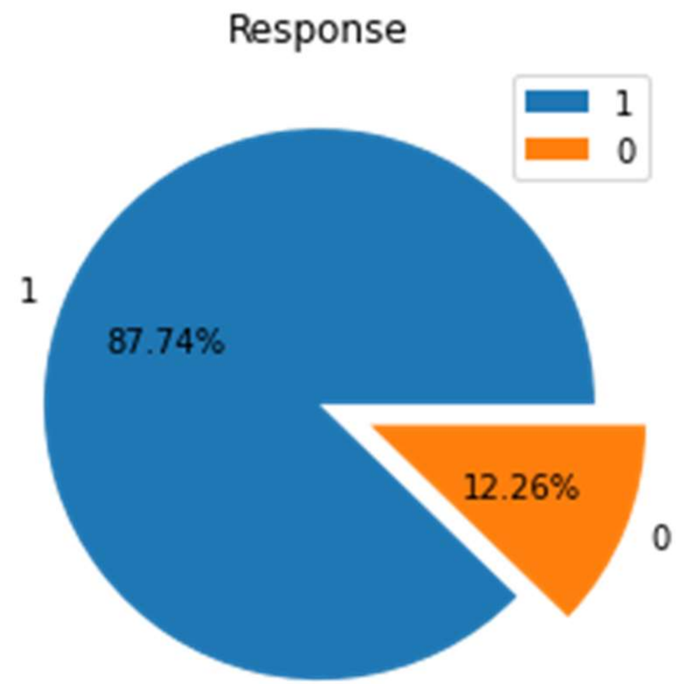
- Vehicle_Damage and Previously_Insured
- Policy_Sales_Channel and Age
- Policy_Sales_Channel and Vehicle_Age



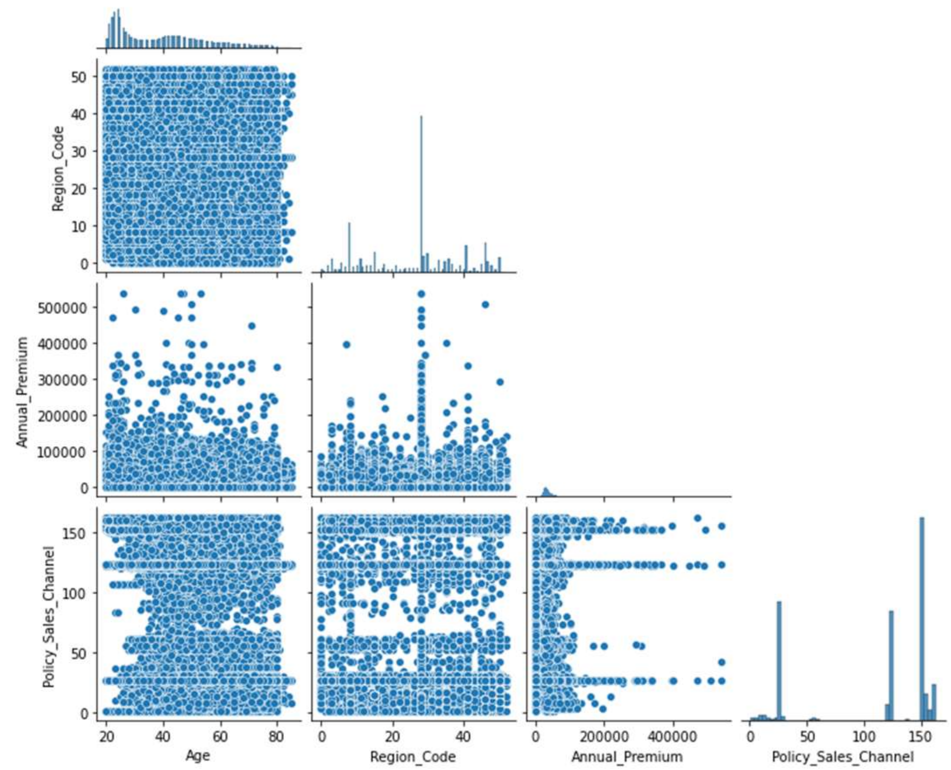
Proportion of Target Variable

The proportion of Response classes in train data set:

- 1: Customer is interested = 87.74%
- 0: Customer is not interested = 12.26%



Pairplot

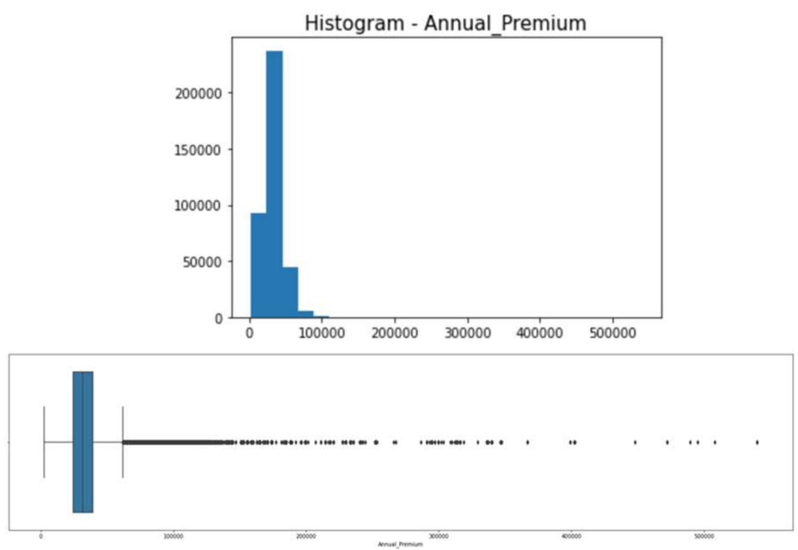


Region_Code
Annual_Premium
Policy_Sales_Channel
Age
Region_Code
Annual_Premium
Policy_Sales_Channel

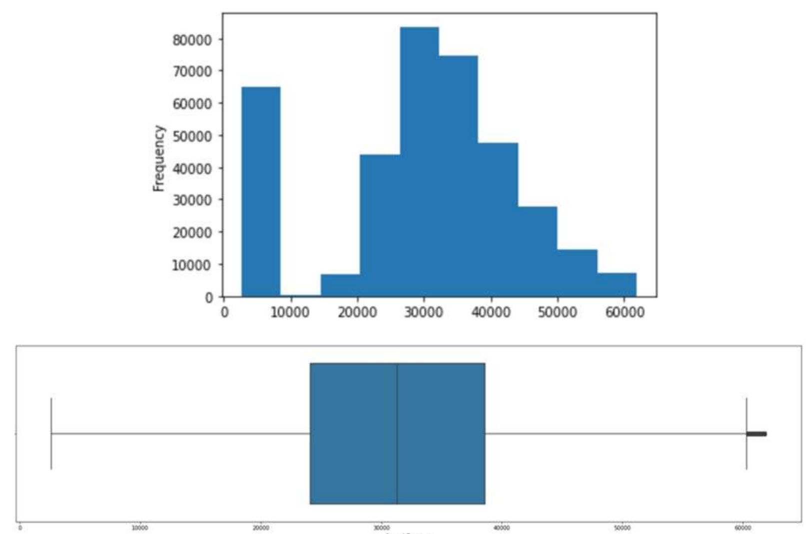
Handling Outliers and Skewness

Annual Premium

(There are a lot of outliers that skews the data)



(After removing 10320 outliers)

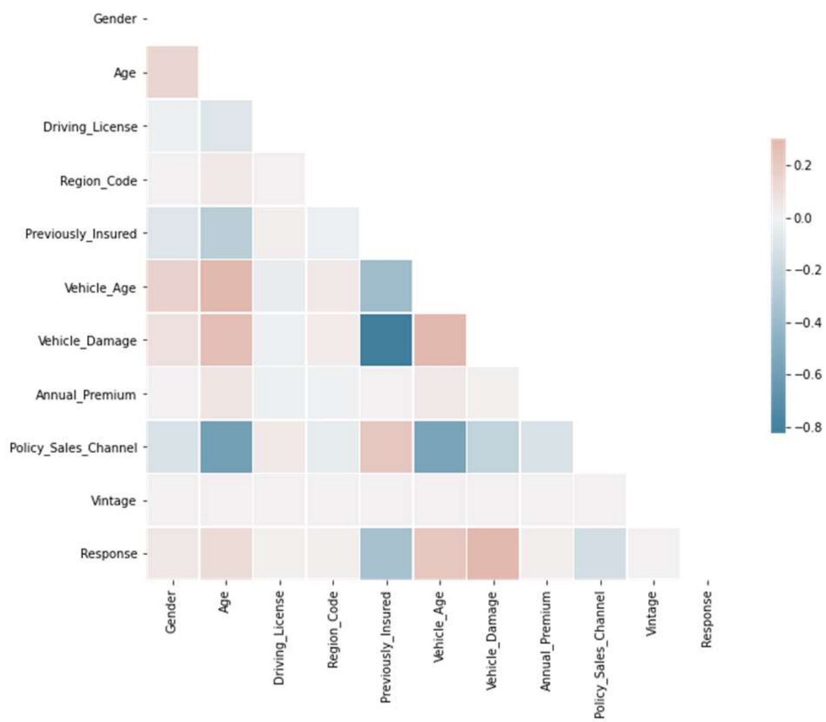


Research Questions

What variables can lead to an increase in insurance premium?

- Looked at correlation between other variables and premium variable.
- Built a Random Forest model to predict Premium.
- Did not find any strong variables that can be big predictors of insurance premium.

Is vehicle damage correlated with any other factors?




```
The RMSE of the training data is 0.07727719041778874
The RMSE of the validation data is 296.45991877967003
```

Correlation Analysis and Linear Regression.

Vehicle damage goes up with:

1. The age of the vehicle
2. Age of the person

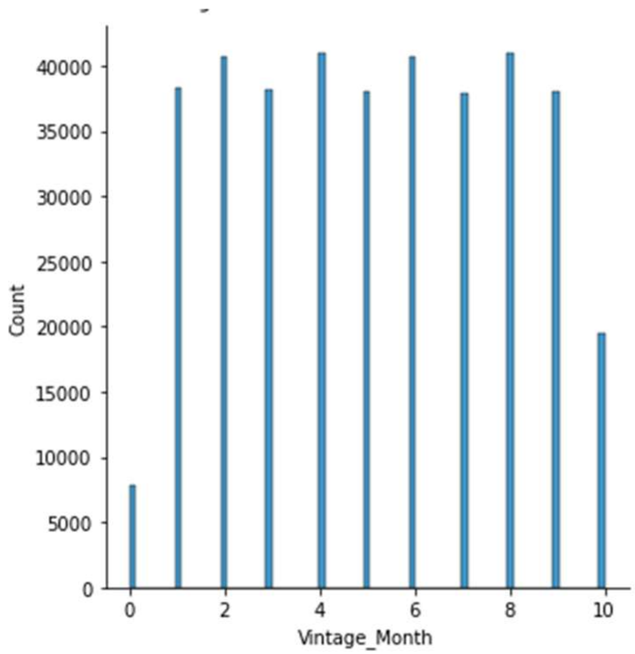
Vehicle damage goes down with:

1. Those who are previously insured

People who are older and have older cars are more at risk of having vehicle damage.

Can We Predict Customer Loyalty ?

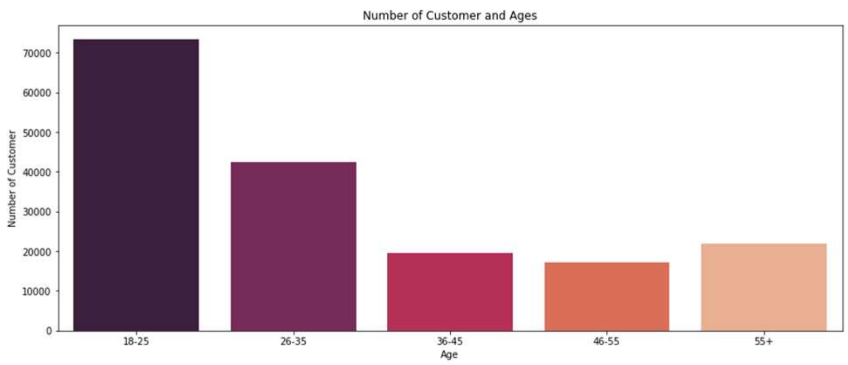
Length of time the customer has been with Insurance company

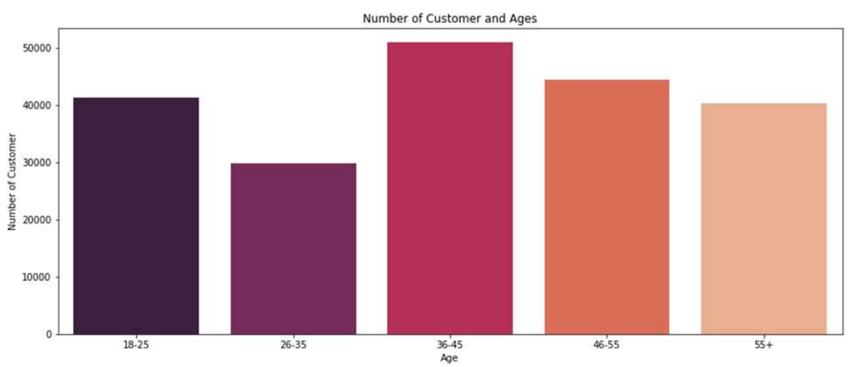


1. The length of loyalty is about the same from the 1st month to the 9th month.
2. With super loyal customers at 10 months being the least common of all.

There is a drop off period around the 9th month. There are 40,920 customers that has been with them for 8 months but only 19507 customers that stay for 10 months.

What type of people have insurance?





People who were not previously insured had:

1. A higher mean age
2. Higher mean vehicle age
3. Higher mean vehicle damage

While those who were previously insured:

1. Had less vehicle damage

Main Research Question

Let’s predict whether health insurance policyholders will also be interested in purchasing vehicle insurance.

Predictive Modeling
&
Evaluation

Classification Algorithms

RandomizedSearchCV

K-fold Cross-Validation
Hyper-parameter Tuning

- Naive Bayes (Gaussian) classifier
- Decision Tree Classifier
- Linear Discriminant Analysis (LDA) Classifier
- Rocchio Classifier
- Random Forest Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- Logistic Regression Classifier

Predictive Modeling & Evaluation

Prepare Data for ML

- Convert / Transform Categorical Variables into dummy variables
- Separate target attribute
- Split the data into train and validation sets using Stratified Sampling
- Standardization of Numerical Data using Min-Max Normalization.

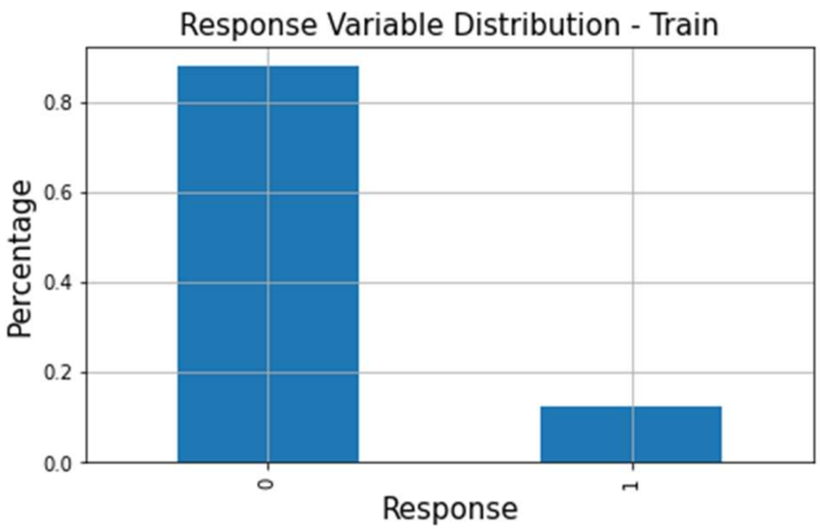
Build, Evaluate, Predict

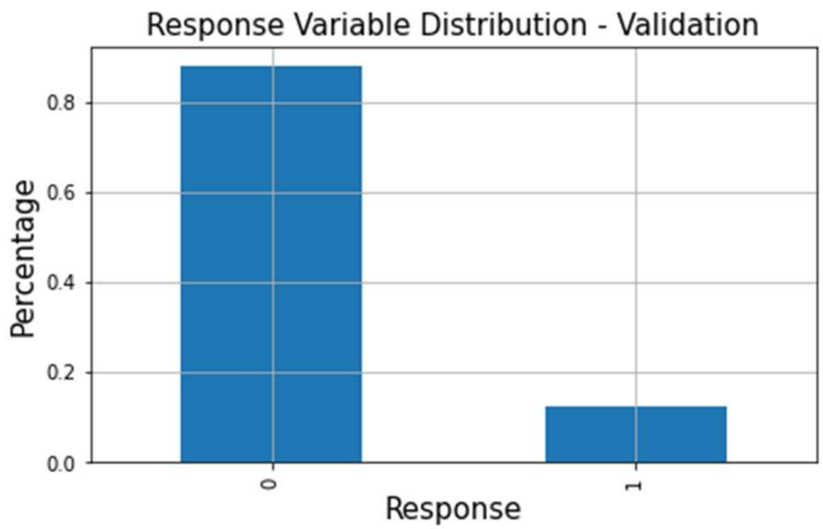
- Build Classification Models
- Calculate ROC_AUC Scores
- Visualize Confusion Matrix
- Visualize Classification Report
- Cross-validation and Hyper-parameter Tuning using RandomizedSearchCV.
- Compare the performance of all models.
- Choose Winner model.

Proportion of Target Variable Classes

Stratified Sampling

Evaluation Metric: ROC_AUC





Performance of All Classification Models

Model	Train ROC_AUC	Test ROC_AUC	Precision	Recall	F1 Score	Support	ROC_AUC Difference
NaiveBayes BaseModel	0.784	0.784	0.903070	0.638923	0.699933	None	0.000
NaiveBayes RSCV	0.784	0.784	0.903070	0.638923	0.699933	None	0.000
DecisionTree BaseModel	1.000	0.599	0.826409	0.821967	0.824145	None	0.401
Rocchio BaseModel	0.577	0.575	0.811701	0.705020	0.747054	None	0.002
RandomForest BaseModel	1.000	0.546	0.823457	0.866220	0.835849	None	0.454
GradientBoost RSCV	0.505	0.504	0.833353	0.877489	0.822452	None	0.001
LinearDiscriminant BaseModel	0.501	0.501	0.814647	0.877201	0.820644	None	0.000
DecisionTree RSCV	0.500	0.500	0.769896	0.877437	0.820156	None	0.000
LinearDiscriminant RSCV	0.500	0.500	0.769896	0.877437	0.820156	None	0.000
RandomForest RSCV	0.500	0.500	0.769896	0.877437	0.820156	None	0.000
AdaBoost BaseModel	0.500	0.500	0.769896	0.877437	0.820156	None	0.000
AdaBoost RSCV	0.500	0.500	0.789678	0.877161	0.820146	None	0.000
GradientBoost BaseModel	0.500	0.500	0.769896	0.877437	0.820156	None	0.000
LogisticRegression BaseModel	0.500	0.500	0.892479	0.877463	0.820220	None	0.000
LogisticRegression RSCV	0.500	0.500	0.892479	0.877463	0.820220	None	0.000

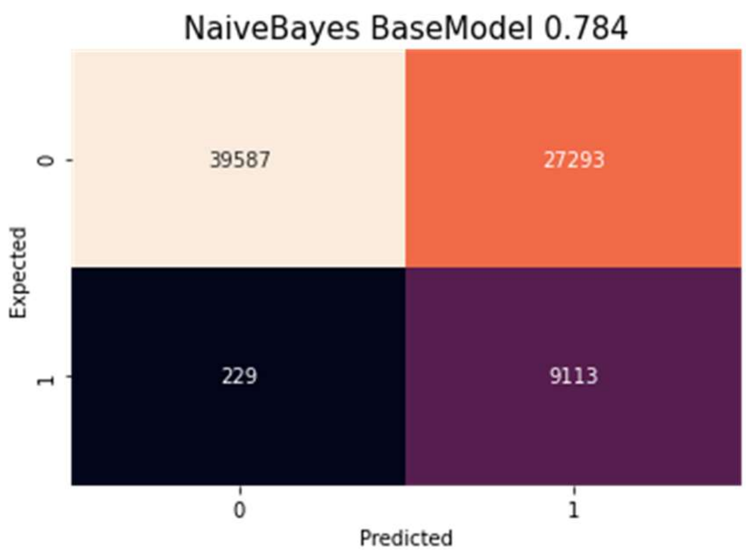
RSCV = RandomizedSearchCV

Winner Model

Naive Bayes (Gaussian) Classifier Base Model

ROC_AUC = 0.784

Run Time = 2 Seconds



	precision	recall	f1-score	support
0	0.99	0.59	0.74	66880
1	0.25	0.98	0.40	9342
accuracy			0.64	76222
macro avg	0.62	0.78	0.57	76222
weighted avg	0.90	0.64	0.70	76222

```
Train ROC_AUC: 0.784
Test ROC_AUC: 0.784
```

Conclusions

- Naive Bayes (Gaussian) Classifier had the highest ROC_AUC score of 0.784 for both training and test (hold-out) sets.
- This model can predict the customers who are interested in vehicle insurance which will ultimately help the company to plan its communication strategy and increase the revenue.
- If we had more time, we could build more models after undersampling the majority class or oversampling the minority class in order to rebalance our dataset.

References

- Machine Learning in Action, by Peter Harrington, Manning Publications, 2012 https://www.manning.com/books/machine-learning-in-action
- Wikipedia: https://www.wikipedia.org/
- Towards Data Science: https://towardsdatascience.com/
- https://scikit-learn.org/
- Python for Data Analysis Book: https://wesmckinney.com/pages/book.html
- KDnuggets: https://www.kdnuggets.com/
- https://www.researchgate.net/profile/P-Pintelas/publication/228084509_Handling_imbalanced_datasets_A_review/links/0c960517fefa59fa6b000000/Handling-imbalanced-datasets-A-review.pdf
- https://www.ijrter.com/papers/volume-3/issue-4/a-review-on-imbalanced-data-handling-using-undersampling-and-oversampling-technique.pdf
- https://link.springer.com/chapter/10.1007/978-981-4585-18-7_2

Thank you, Questions? ◡̈