Health Insurance Cross Sell Prediction

IST 718 Group 6
Project Report

Aatish Suman
Adhessh Phadnis
Sharvil Turbadkar
Snehal Ghatpande

Table of Contents

1. Abstract 3
1.1 Project Overview 3
1.2 Predictions 3
1.3 Inferences 3
1.4 Conclusion summary 3
2. Data Collection, Cleaning and Exploration 4
2.1 Description of dataset 4
2.2 Preliminary data exploration: 5
2.3 Feature engineering 7
3. Methodology 7
3.1 Feature selection 8
4. Models 9
4.1 Naive Bayes 9
4.2 SVM 10
4.3 Logistic Regression 10
4.4 Random Forest 11
4.5 Gradient boosting 13
4.6 Neural Networks - Multilayer Perceptron 14
5. Conclusion 14

1. Abstract

1.1 Project Overview

The objective of the project is to build a model to predict whether the policyholders from the past year will also be interested in purchasing vehicle insurance provided by the company. Building the prediction model would be extremely helpful for the company. It can aid the company to plan its communication strategy to reach out to those customers and optimize its business model and revenue. To predict whether the customer would be interested in Vehicle insurance we are using the data about the demographics (gender, age, region code type), vehicles (age, damage), policy (premium, sourcing channel), etc.

1.2 Predictions

We used the Naïve Bayes-, SVM-, Logistic regression-, Random forests-, Gradient-boosted trees- and ANN-based classifiers to predict whether a current policyholder would be interest in purchasing vehicle insurance from the company using customer demographics, vehicle properties, and policy.

1.3 Inferences

- Customers who do not have vehicle insurance are predicted to buy the vehicle insurance from the company.
- Customers who have reported vehicle damage in the past are predicted to buy the vehicle insurance from the company.
- Customers in the age group of 30-55 are predicted to buy the vehicle insurance from the company.
- Policy sales channel 149 can be the most effective way of reaching out to the customer in order to sell the vehicle insurance plan.
- Customers who have to pay higher annual premium plans are more likely to buy vehicle plans

1.4 Conclusion summary

We achieved the best PR AUC of 0.6696 and accuracy of 78.3 for the Random Forest model. Previously_insured (whether the customer already has a vehicle insurance or not), Vehicle_damage (whether the customer has damaged their vehicle or not), age of the customer, Vintage (how long the customer has been associated with the company) and Policy_Sales_Channel (specific types of policies) are found to be the most important predictors.

2. Data Collection, Cleaning and Exploration

2.1 Description of dataset

The dataset was obtained from Kaggle - https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction.

The training data set consists of 381,109 rows and 12 columns while the test data set consists of 127,037 and 11 columns. Only the training dataset was used for analysis since the test set did not have the target variable.

Variable	Description
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 i.e. 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

Table containing the list of the variables along with their descriptions

2.2 Preliminary data exploration:

Target variable – The target variable (Response), representing whether the customer is interested in vehicle insurance or not, is heavily imbalanced as can be seen from the bar plot below.

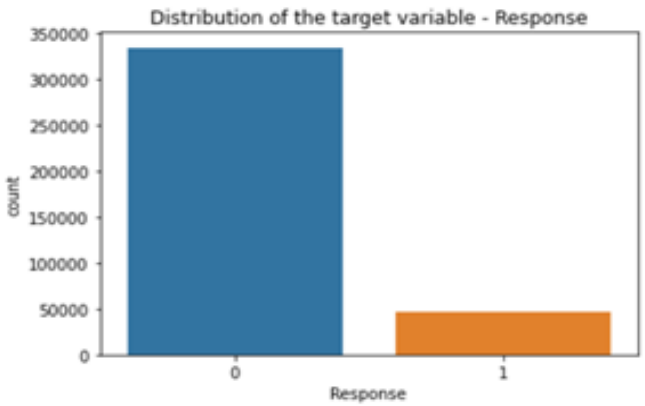


Figure showing the distribution of the target variable

Missing values – The min, mean and max values for each numerical variable are comparable, which reduces the possibility of erroneous data. The data does not have null values. The scales of the variables are different, and scaling might be required for certain algorithms and to assess variable importance.

summary	Age	Region_Code	Annual_Premium	Policy_Sales_Channel	Vintage
count	381109	381109	381109	381109	381109
mean	38.822583565331705	26.388807401557035	30564.389581458323	112.03429465061177	154.34739667654136
stddev	15.511611018095289	13.22988802578849	17213.15505698001	54.20399477485634	83.67130362658735
min	20	0.0	2630.0	1.0	10
max	85	52.0	540165.0	163.0	299

Table showing the summary of the numerical variables

Variable correlation – There doesn't seem to be any correlation between the variables Age, Region_Code, Annual_Premium, Policy_Sales_Channel and Vintage as can be seen from the pair plots below.

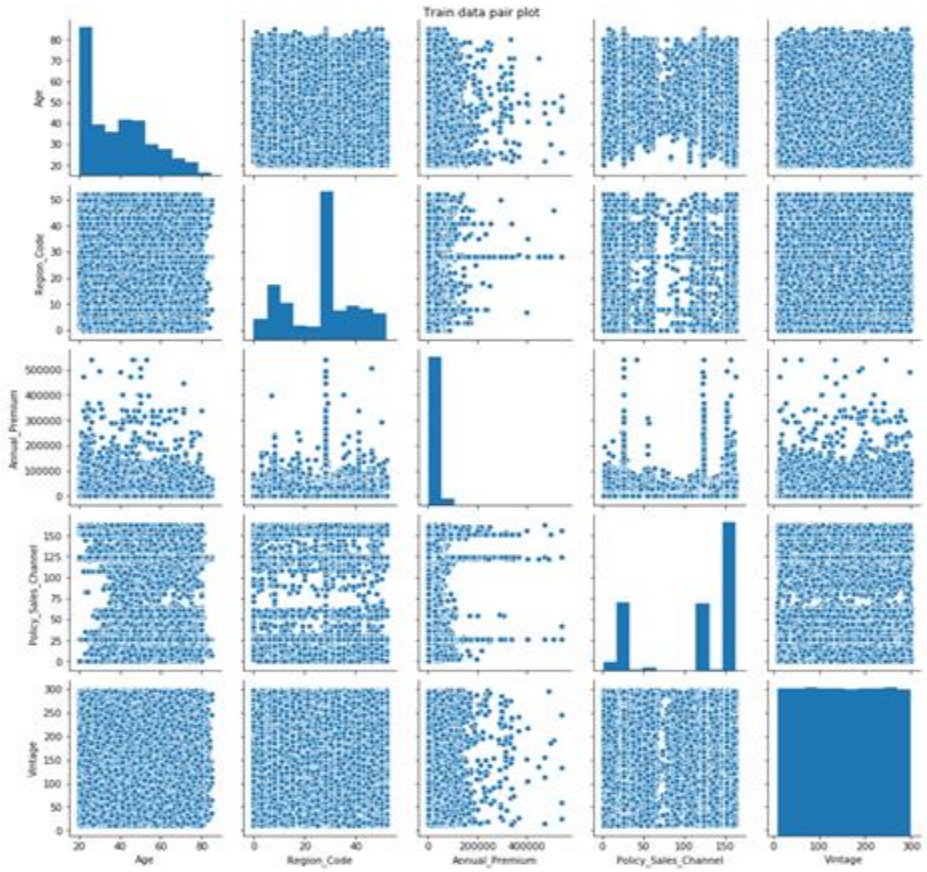


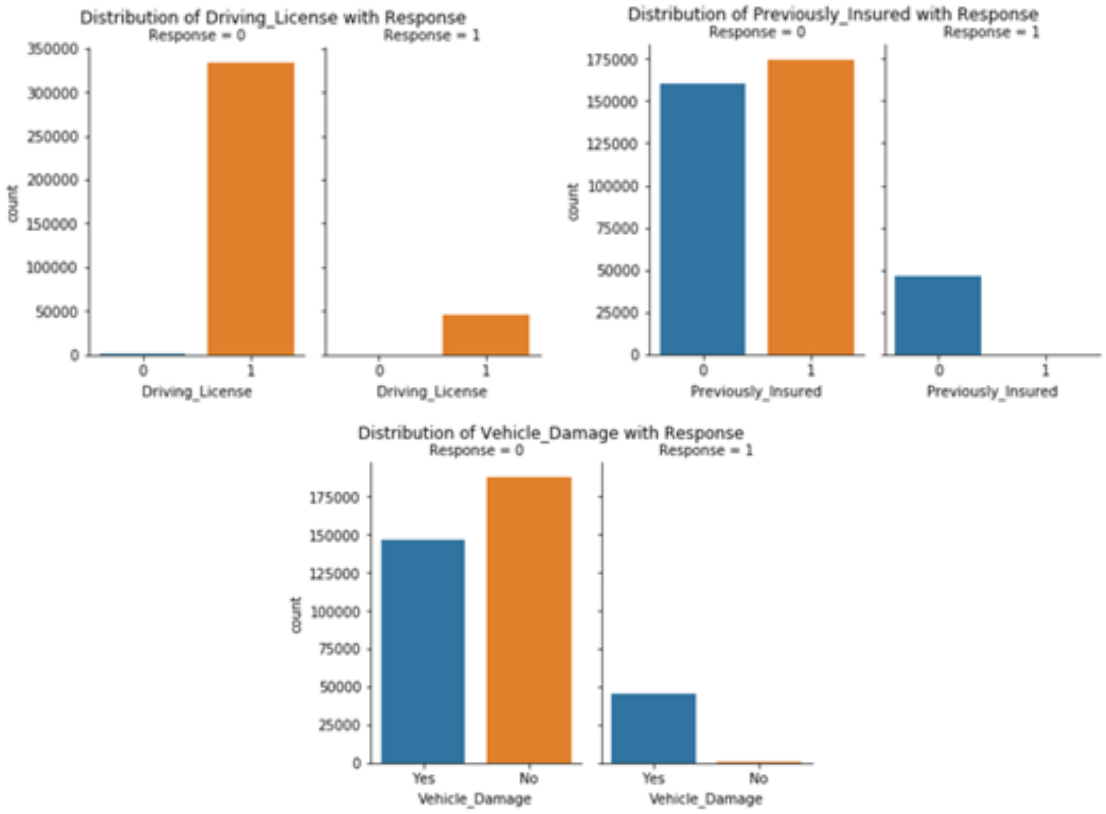
Figure showing the pair-plots of the numerical variables

Interesting finding – All interested customers (Response=1) have a driving license (Driving_License=1) and already have a vehicle insurance (Previously_Insured=1), and almost all of them have damaged their vehicle in the past (Vehicle_Damaged=Yes) (shown in the categorical plot below). It shows that the 3 variables have high predictive power.Elderly people are most likely to buy the vehicle plan over younger people. Thus, we can make an inference saying elderly people are most likely to damage their vehicle and get a vehicle plan than younger people. Also elder people are more likely to pay more for their vehicle plan to avail the most of the services

PremiumCategory	0	1	Ratio
High	112614	18591	0.16508604613991157
Low	100066	12978	0.1296944016948814
Medium	121719	15141	0.12439306928252779

AgeCat	0	1	Ratio
Old	81646	13161	0.16119589447125396
Young Adult	80187	7820	0.09752204222629603
Elder	82010	22416	0.2733325204243385
Adolescent	90556	3313	0.03658509651486373

Tables showing ratio of Response by age and annual premium



Figures showing the distribution of *Driving_License* (top left), *Previously_Insured* (top right) and *Vehicle_Damage* (bottom) across the target variable

2.3 Feature engineering

For all the models, the categorical variables *Gender*, *Vehicle_Age* and *Vehicle_Damage* were converted to numerical form by using the StringIndexer class. Vehicle_Age and Vehicle_Damage are ordinal variables and the StringIndexer class makes sure that the rank of the values is preserved. *Gender* contains only 2 values and therefore, any method to convert it to numerical form is sufficient. The categorical variables *Region_Code* and *Policy_Sales_Channel* are already present in numerical form but using one-hot encoding for certain algorithms like logistic regression provided better results.

3. Methodology

The steps of the methodology at a high-level as shown in the figure below were –

1. Using linear regression and random forest to understand feature significance.

2. Using the important features for training the models using Naïve Bayes, SVM, Logistic regression, Decision tree, Random forest, Gradient boosted tree and ANN based algorithms.
3. Using the best model to make inferences about the dataset.

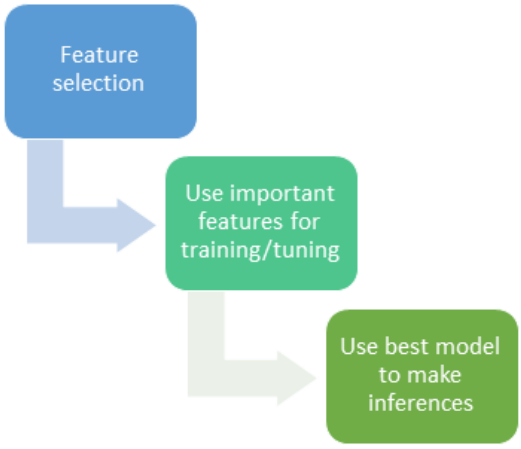
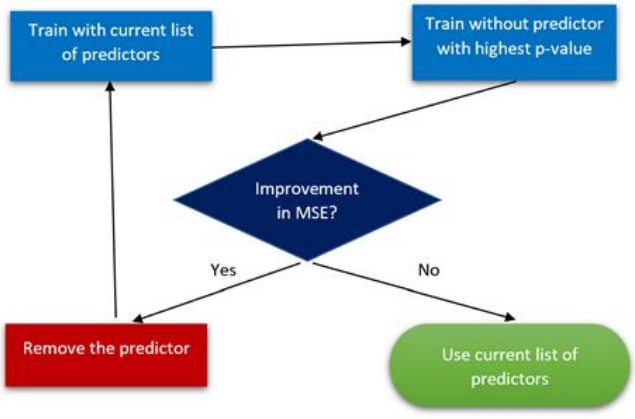


Diagram showing the steps of the methodology at a high-level

3.1 Feature selection

Two methods were used to understand feature significance – a linear regression stepwise selection method based on p-value and MSE, and the feature importance property of the random forest model. The steps of the first method as shown in the figure below were -

a) Starting with all predictors measure the MSE.
b) Sort the predictors by p-values and remove the one with the highest p-value.
c) Measure the MSE and check if there's an improvement.
    d) If yes, remove the predictor, and if no, use the current list of predictors.



Flow diagram showing the steps of the linear regression stepwise selection method

The final list of predictors obtained from this method were – *Vehicle_Damage, Vehicle_Age, Previously_Insured, Driving_License, Policy_Sales_Channel, Annual_Premium* and *Age.*

The table below shows the predictors with the corresponding p-values as obtained in the first iteration of the linear regression method and the corresponding values of the feature importances as obtained by a quick and dirty implementation of the random forest algorithm. The predictive models were trained using different combinations of the features using all the information obtained about the features during this analysis. The details of the models are provided in the next section.

Variable	Linear regression p-value	Random forest feature importance
Vehicle_Damage	0.000000e+00	0.419990
Age	0.000000e+00	0.280382
Policy_Sales_Channel	0.000000e+00	0.140696
Vehicle_Age	0.000000e+00	0.079464
Previously_Insured	0.000000e+00	0.075979
Region_Code	2.226673e-01	0.003271
Vintage	9.445091e-01	0.000178
Annual_Premium	0.000000e+00	0.000040
Gender	1.432632e-12	0.000000
Driving_License	0.000000e+00	0.000000

Table showing the list of variables and their corresponding p-values and feature importances

4. Models

4.1 Naive Bayes

Naïve Bayes (NB) classifiers are a family of probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. Although this assumption does not hold true in most cases, they are frequently used to create a quick baseline model. The list of features obtained from the stepwise selection method were used to train a Multinomial Naive Bayes model and grid search was used to tune the smoothing parameter. Since the dataset was highly imbalanced, the area under the PR curve was used along with the validation accuracy to measure the performance of the model. The best value for the area under the PR curve was 0.1602 and that for the validation accuracy was 0.6809. The

validation accuracy was less than the majority vote baseline of 0.8774 and the model was therefore not used for further analysis.

4.2 SVM

The list of features obtained from the stepwise selection method were used to train a Linear SVM model and grid search was used to tune the maximum iterations and the regularization parameter. Since the dataset was highly imbalanced, the area under the PR curve was used along with the validation accuracy to measure the performance of the model. The best value for the area under the PR curve was 0.2948 and that for the validation accuracy was 0.8781. The area under the PR curve was significantly higher than that of the NB model but the validation accuracy was almost the same as the majority vote baseline. Other advanced models were tried and this model was not used for making any inferences.

4.3 Logistic Regression

Logistic regression is a classification algorithm used for binomial/multinomial classification problems. The algorithm is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, which in our case is whether or not the customer will be interested in buying the vehicle insurance given he already has the health insurance from the company.

Phase 1:

We tried to implement different angles to model the data using Logistic regression. For basic preprocessing we indexed the ordinal and nominal variables viz. Gender, Vehicle Age and Vehicle Damage. From analyzing the categorical variables Region code and Policy sales channel, it could be observed that these variables had a large number of categories. From the feature selection methodology using Linear Regression, it was observed that region code had p-value > 0.05 making it statistically insignificant. Hence, we decided to drop this variable thus solving the problem of high categories. After one hot encoding the policy sales channel variable, we fitted a basic LR model on the preprocessed data which yielded --0.33 PR AUC value and 87.51% accuracy. To reduce the exposure for overfitting we then used cross validation for tuning the regularization parameters Viz. elastic net and regression. However, not much difference was noted in the evaluation metric values. Fitting the best model extracted from cross validation on validation data yielded approximately the same values.

Phase 2:

One of the key problems with the dataset is the class imbalance in the response variable. Such a dataset can cause your model to blindly predict dominant class since it can achieve good accuracy anyway. We decided to go for Oversampling method which by its name oversamples the minority class until the ratio of majority and minority class labels is approximately equal. The oversampling is done on the training data set from the train test split. The model is then evaluated on the validation set from the original data. This process is aimed at generalizing the model and taking the majority class bias out of the equation. Further preprocessing of the data is the same as Phase 1 except that the variable Vintage is dropped from the data and region_code is kept in by one hot

encoding. Instead of using age as a continuous variable we tried to bin it into three categories which would define the age as young, adult and old. Finally, using cross validation we achieve 0.3265 as PR AUC and 69.37% accuracy. This model can be said to be more generalized than the phase 1 model since it is trained on unbiased data. By extracting the coefficients of the LR equation we arrived at the result that customers with certain Policy_Sales_Channel have the highest effect on the odds of them buying the Vehicle insurance.

4.4 Random Forest

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of bagging is to aggregate bootstrapped datasets from the models (bootstrapped aggregation) and combining the learning models to increase the overall result.
Categorical variables like Vehicle Age, Vehicle Damage, and Gender were discretized after which columns were imputed using log transformation as many predictor variables had different units of measurements with Annual Premium having very high values. To ensure data was consistent we standardized using log imputation.

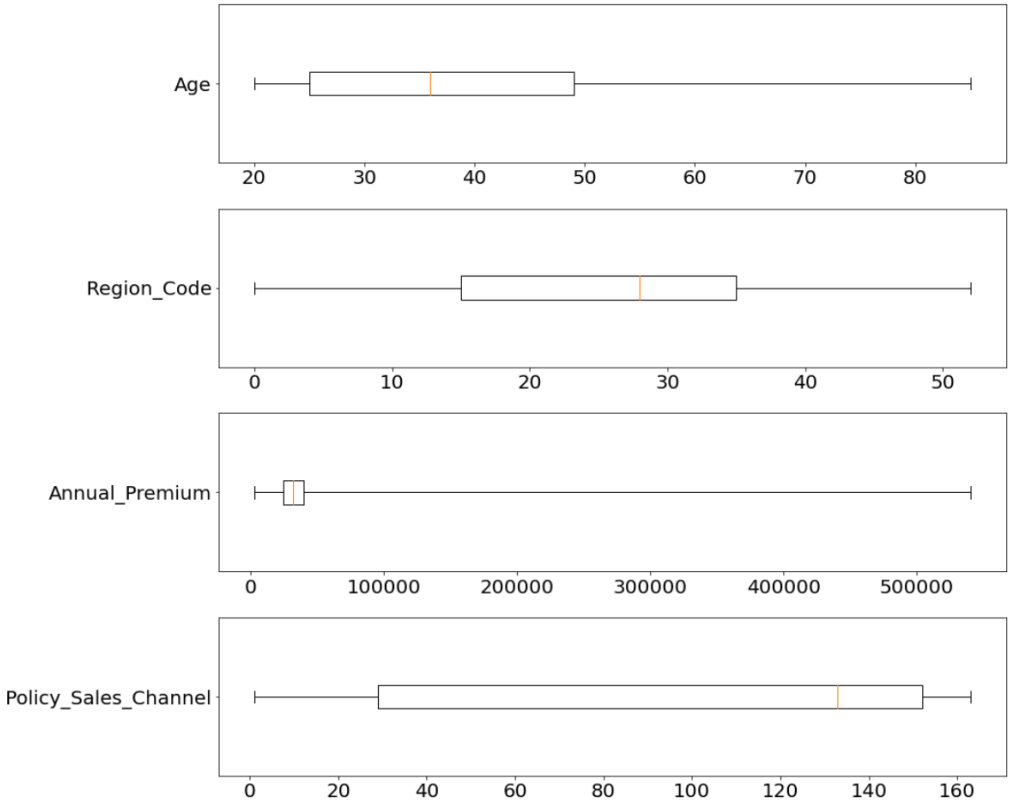


Diagram showing boxplots for all continuous variables

After visualizing the boxplots, it was observed that variables like age and annual premium are heavily left-skewed. To ensure the effect of outliers is not prominent we winsorized the data on the 90th and 5th percentile to not lose any data and ensure outliers do not affect the outcome.
Hyperparameter tuning: The hyperparameters tuned were maxDepth, maxBins and numTrees. The maxDepth parameter was selected because allowing the trees to grow to a greater depth leads to higher variance. The numTrees parameter allows us to reduce variance by increasing the number of trees. The maxBins allow us to decide the maximum number of splits to be allowed while discretizing the categorical variables. Increasing maxBins allows the algorithm to consider more split candidates and thus increase variance.
First Approach:
After using a vector assembler, we fitted a random forest model and used grid search to fit the model across various hyper parameters and get the best metrics. To ensure our model performs has a lower generalization error and performs its best in a testing environment cross-validation was performed The classifier is trained on many samples where we know the response, but our goal is to predict how the classifier predicts samples for a new plan it has never seen before. We test our classifier using 3-fold cross-validation to train data on two folds.
We fit the cross validated model on the train data and evaluated its performance on the validation data with an accuracy of 87.83% and a PR-AUC Score of 0.35.
Second Approach:

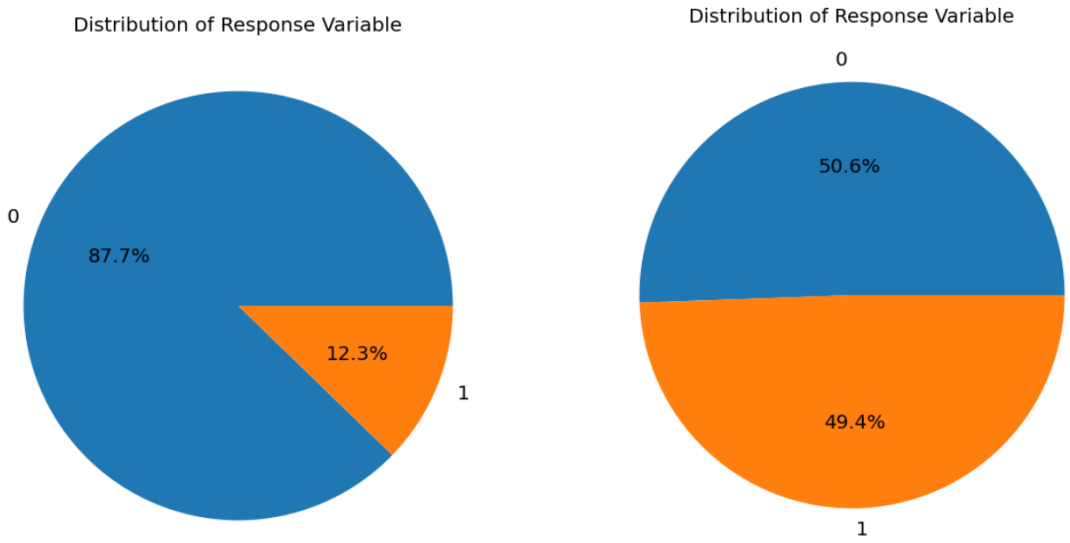


Diagram showing distribution of Response before and after oversampling

We tried to address the problem of class imbalance by performing oversampling on the minority data to ensure high accuracy is not a result of the majority vote baseline. The oversampling is done on the training data set from the oversampled data and is then evaluated on the .In order to ensure random forest does not treat categorical variables as continuous variables for split, we use a vector indexer for the model to identify and index categorical variables. This information is stored as metadata and improves model quality significantly.
After using a vector indexer after the vector assembler, we perform fit on the oversampled cross-validated model and test the results original validation set from the original data.

We get an accuracy of 78.30% and a PR-AUC score of 0.669.
This model is a fair and generalized representation of how the model will perform on the testing data. Also, there is a significant increase in the PR-AUC score.

4.5 Gradient boosting

Gradient boosting builds an ensemble of trees one-by-one. It is a sequential algorithm that keeps adding predictors to the ensemble model with each one correcting its predecessor
The next decision tree tries to cover the discrepancy between the target function and the current ensemble prediction of whether the customer would buy the health plan and thus re-compute the residual and feed-forward its output to compute the next round of predictions.
The tuned params for this model are -
Max Depth:16, Maximum number of Iterations: 20
Before using the vector, indexes and oversampling the metrics obtained were as follows:
Accuracy:87.64% **PR AUC Score**:0.3557
After using the vector indexer for indexing columns and oversampling the minority class the metrics obtained were as follows.
Accuracy:77.94% **PR AUC Score**:0.6156
The accuracy we now obtain is a generalized and more accurate representation of well will the model perform. The PR AUC Score also shoots up after predicting the original cross validated data over the oversampled data.
Feature Importance for Random Forest

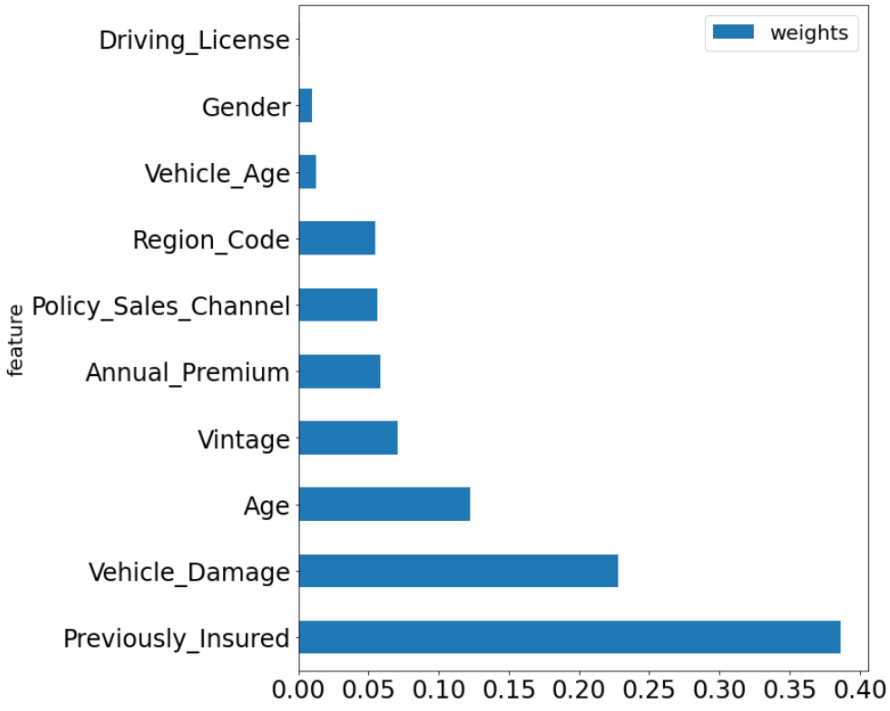


Diagram showing feature importance of all predictor variables for Response =1

As we can see Previously insured Vehicle Damage and Age and Vintage become the most important features.
Inferences: Since Previously Insured is the most important feature according to random forest and Gradient Boosting, we infer that customers who do not have vehicle insurance are the ones most likely to buy the vehicle plan. Customers whose vehicle has been damaged are most likely to buy the vehicle plan from the customer.

4.6 Neural Networks - Multilayer Perceptron

We used Multilayer Perceptron, which is the classical type of neural network. It consists of one or more layers of neurons. Significant features obtained through feature importance like Vehicle_Damage, Vehicle_Age, Previously_Insured, Driving_License, Policy_Sales_Channel, Annual_Premium and Age were fed to the input layer. The dataset that we were working on was not too complex, so we decided to try and use a small number of hidden layers which provided the levels of abstraction. We made predictions on the output layer, in this case it was based on the **Response** variable.

We started out with no hidden layers in the model and used the AUC metric to score it. We added more layers and changed the number of neurons in the hidden layers to find a good value for the AUC metric. Finally we used grid search and cross validation for hyper tuning the various parameters like the layers and the neurons in the layers, the block size for stacking input data, tolerance, learning rate (step size) that controlled how much to change the model in response to the estimated error each time the model weights were updated in order to find the best model. Using the best model, we were able to achieve 0.4990 as PR AUC and 87.74% accuracy.

Vehicle_Damage	Vehicle_Age	Previously_Insured	Driving_License	Policy_Sales_Channel	Annual_Premium	Age	prediction
1.0	2.0	0	1	26.0	40454.0	44	1
0.0	1.0	0	1	26.0	33536.0	76	0
1.0	2.0	0	1	26.0	38294.0	47	1
0.0	0.0	1	1	152.0	28619.0	21	0
0.0	0.0	1	1	152.0	27496.0	29	0

By looking at the prediction column, we inferred that the damage that has happened to a vehicle and the age of the vehicle would play a key role in the chance that a customer would buy vehicle insurance. The company can thus use this data and target customers with old and damaged vehicles to cross sell vehicle insurance to them.

5. Conclusion

In this project we explored detailed analysis and modeling of Health insurance cross selling data using sophisticated data science techniques using big data infrastructure to predict whether a customer is/will be willing to buy vehicle insurance given he has health insurance

from the company. We were able to extract customer characteristics based on results given by statistical and machine learning algorithms which might aid the company to strategically isolate customers based on personal characteristics like age or vehicle_damage state etc. create targeted marketing strategies, advocate different sales policies which are proven statistically important for the customers to be willing to buy the vehicle insurance.

We successfully implemented distance based (SVM), ensemble (Random Forest and GBT) , probabilistic (Logistic Regression, Naïve bayes) and advanced AI model (ANN) with decent accuracy. Addressing the problem of class imbalance, we implemented oversampling techniques to train more generalized models. We achieved the best PR AUC of 0.6696 and accuracy of 78.3 for the Random Forest model. The table below summarizes the scores for all the models.

Model Name	**PR AUC**	**Accuracy (%)**	**Oversampling (Yes/No)**
Random Forest	0.6696	78.3	Yes
ANN	0.4990	87.74	No
GBT	0.6156	77.94	Yes
LR	0.3265	69.37	Yes
SVM	0.2948	87.81	No
NB	0.1602	68.09	No

Model Comparison

Nonetheless, the feature importance results from other models provide valuable information which can also be taken into consideration from a more holistic business perspective. The specific inferences that the company can look at are:

- Customers who do not have vehicle insurance are predicted to buy the vehicle insurance from the company.
- Customers who have reported vehicle damage in the past are predicted to buy the vehicle insurance from the company.
- Customers in the age group of 30-55 are predicted to buy the vehicle insurance from the company.
- Policy sales channel 149 can be the most effective way of reaching out to the customer in order to sell the vehicle insurance plan.
- Customers who need to pay a higher annual premium are more likely to buy the vehicle plan

The future scope for the project is to research and add more variables specifically relating to the health insurance policy, financial characteristics like income, job type etc. The models can then be trained and tested on this data for better or worse prediction accuracy till the model becomes reliable enough.