Roma Coffin
Data 1030 Final Project
GitHub Repository

# Health Insurance Cross Sell Prediction

## 1.0 - Introduction:

For my machine learning project, I analyzed a data set that is aimed to see whether or not someone will be interested in purchasing vehicle insurance. The goal is for a health insurance provider to leverage their own customer's data from past years to see if they would also be inclined to buy car insurance from them. The Kaggle dataset used is for a current competition to build a model to predict whether the policyholders (customers) will be interested in vehicle insurance provided by the company. The test set provided in Kaggle is used for the contest ranking, I ignored that dataset and focused on the training data, which resulted in splitting the given training data into train, validation, and test, so I could use different random seeds for splitting.

The outcome for this problem is discrete because we are trying to answer yes or no (1 or 0) if someone is interested in purchasing vehicle insurance and thus it is a classification problem. The target variable is the customer's response, if they are interested in buying car insurance. The dataset has 381,109 datapoints. See the table below for a short description on each of the 12 features.

| Feature | Numerical/Categorical | Description |
|---|---|---|
| Id | NA | Unique ID of each customer |
| Gender | Categorical | Male or Female |
| Age | Numerical | Age of customer |
| Driving License | Categorical-0,1 | If they have a driver's license (0) or not (1) |
| Region Code | Numerical | Region customer is from (0-52) |
| Previously Insured | Categorical-0,1 | If customer already has insurance (1) or not (0) |
| Vehicle Age | Categorical | <1 year, 1-2 years, 2+ years |
| Vehicle Damage | Categorical-0,1 | If car has been damaged before (1) or not (0) |
| Annual Premium | Numerical | Annual Premium Amount in Rs. |
| Policy Sales Channel | Numerical | Customer outreach channel (1-163) |
| Vintage | Numerical | Time as a customer (10-299) |
| Response/Target | Categorical-0,1 | Customer is Interested (1) Not Interested (0) |

## 2.0 - Exploratory Data Analysis:

A thorough analysis was performed on each column in the dataset. A summary of the most important findings can be seen below.

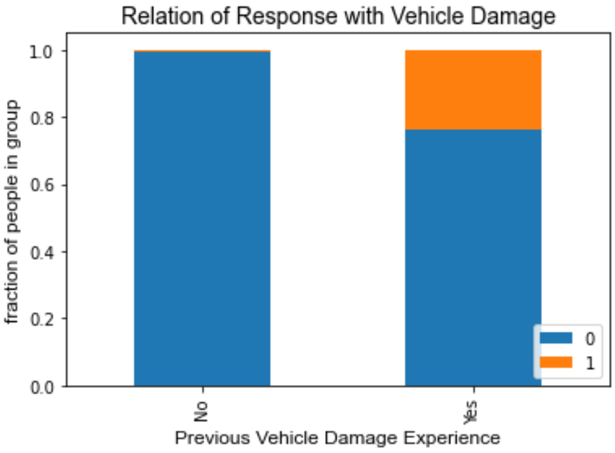

**Figure 1:** *This stacked bar plot shows that people with past vehicle damage are more likely to be interested in purchasing vehicle insurance while people with no past vehicle damage are less likely to be interested in purchasing vehicle insurance.*

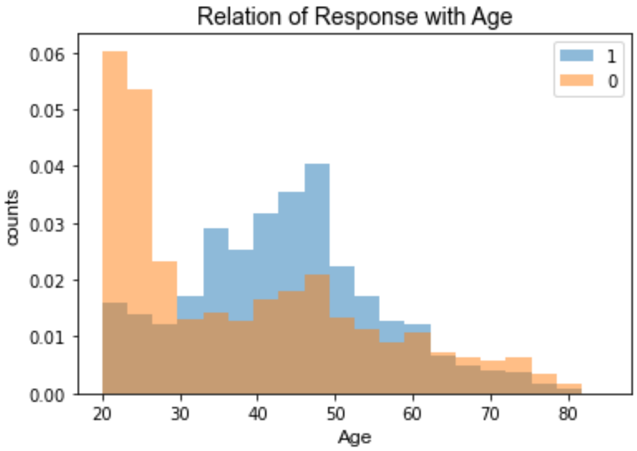

**Figure 2:** *This category specific histogram shows that people between the ages of 35 to 50 are most likely to be interested in purchasing vehicle insurance while people between the ages of 20 to 30 are less likely to be interested in purchasing vehicle insurance.*

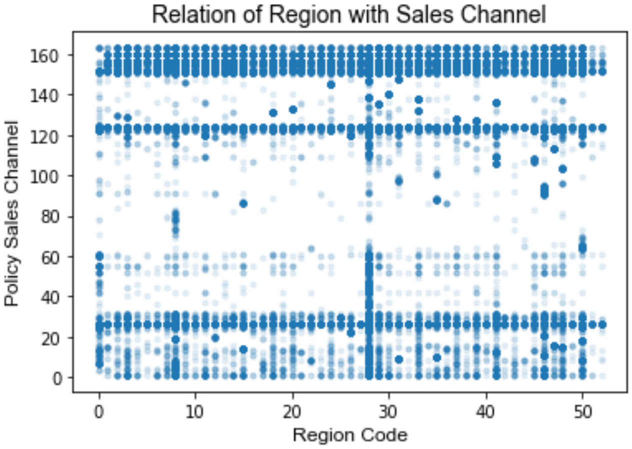

**Figure 3:** *The scatterplot for Region Code and Policy Sales Channel shows us that Region 28 and Policy Sales Channel 28, 125, and 150 have high volumes. By also looking at the individual scatterplots for both of these features by Annual Premium we can confirm these findings. Both features show us if certain locations have better sales with certain distribution channels*

### 3.0 - Methods:

After performing EDA on the data, the data was then preprocessed. There were no missing values to address. There are 10 features in the dataset, both Id and Vintage features were removed because no impact to response variable was seen. The data was preprocessed inside the machine learning pipeline. See below on why a specific preprocessor was chosen for each feature.

| Feature | Numerical/Categorical | Preprocessor | Reason |
|---|---|---|---|
| Gender | Categorical | OneHotEncoder | Categorical can't be ordered |

| | | | |
|---|---|---|---|
| Age | Numerical | StandardScaler | Numerical reasonable bounds |
| Driving License | Categorical-0,1 | OneHotEncoder | Categorical can't be ordered |
| Region Code | Numerical | StandardScaler | Numerical |
| Previously Insured | Categorical-0,1 | OneHotEncoder | Categorical can't be ordered |
| Vehicle Age | Categorical | OrdinalEncoder | Categorical can be ordered |
| Vehicle Damage | Categorical-0,1 | OneHotEncoder | Categorical can't be ordered |
| Annual Premium | Numerical | StandardScaler | Numerical, tailed distribution |
| Policy Sales Channel | Numerical | StandardScaler | Default |
| Response/Target | Categorical-0,1 | OneHotEncoder | Default |

After preprocessing the data, the next step in the machine learning pipeline is to split the data appropriately. The data is independent and identically distributed because all of the samples seem to stem from the same generative process which assumes that it has no memory of past generated samples. The data does not have a distinct group structure and it is not time series data because it is not indexed over a specific time period order. The data was split into training and other data. Then in the machine learning pipeline, K-Fold Cross Validation was performed so that the best hyperparameters and models can be found by training, cross-validating, and testing the data.

Three machine learning algorithms were selected:

- Logistic Regression-simple and easy to construe, captures non-linear relationships, and provides smooth predictions
- Random Forest Classifier- handles imbalanced and large datasets well and minimal issues with overfitting
- Support Vector Machine-performs well in high dimensions, captures non-linear relationships, downfall is that it can be slow to implement especially with large datasets

Please see the table below for the parameters which were tuned, and the values tried for each machine learning algorithm.

| **ML Algorithm** | **1-Parameter** | **1-Values** | **2-Parameter** | **2-Values** |
|---|---|---|---|---|
| LR | C | logspace(-2, 2, num=8) | NA | NA |
| RF | Max Depth | 1, 3, 10 | Max Features | .5-1 |
| SVM | C | logspace(-3, 4, num=8) | Gamma | logspace(-3, 4, num=8) |

Once we the ideal model is found, then the next step is to see how well it does in the test data. An evaluation metric is used to help decide how it performs in the test set. Since the objective is to predict as accurately as possible, accuracy score seems ideal, but the data is quite imbalanced. ROC_AUC score also looks like it could be a good choice because this is a binary classification problem but again the data is quite imbalanced. Thus, F-1 score is a better metric because it measures inaccurately predicted cases more appropriately. In the end, accuracy score was picked as a metric because it is what is being asked by the insurance company for the original competition. Additionally, to avoid the risk of uncertainties due to splitting and non-deterministic machine learning models (random forest classifier), various random states of K-folds CV are looped thru to calculate both the mean and standard deviation of the looped thru outcomes.

Additionally, feature importance analyses were completed for each machine learning algorithm to help grasp how the features in the model contributes to the model prediction. For logistic regression, random forest classifier and support vector machines perturbation was performed. SHAP, helps to show the local feature importance by using a ranking methodology for a specific row in the data, due to the limited impact of the features on a global scale, a more local view using SHAP did not seem to be very powerful to do.  Lastly, 2x2 confusion matrices were developed for each algorithm to help measure effectiveness of the model. The confusion matrix uses the random state and afterwards is normalized by row to attain true positives, false positives, false negatives, and true negatives. It is important for the company to know whether a customer is or is not interested in vehicle insurance for its communication and marketing strategy. The company can then influence and reach out to customers accordingly, which will then allow for an improved sales model and more profit.

### 4.0 – Results:

The baseline accuracy score is .877, which reflects that if we guess that someone is not interested in purchasing vehicle insurance, we have about an 88% chance of being right. Once applying our machine learning pipeline, the logistic regression algorithm achieves an accuracy score of 86.7%b which is below the baseline accuracy score of .877, the random forest algorithm achieves an accuracy score of 89.2%, and the support vector classifier algorithm also achieves an accuracy score of 89.2%. These machine learning algorithms do not reflect a significantly improved accuracy score when compared to the baseline model.

By calculating permutation importances, I attempted to help to show the feature importance of various features by using a ranking methodology. There was basically no difference which showed that the features were not important. The figure below reflects that there is no impact from these features.

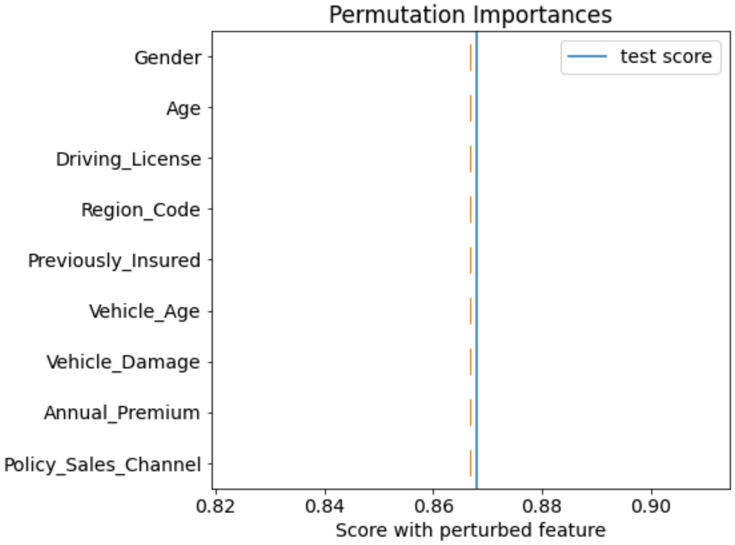

**Figure 4:** *The Permutation Importances boxplot for the Random Forest Algorithm shows us that there is basically no impact from the features.*

Please see the table below for the mean and standard deviation accuracy score for the three different machine learning algorithms. These results show that the random forest and support vector classifier have the best mean score. Logistic regression has the lowest standard deviation which reflects that it is the steadiest in classifying insurance interest among customers for this dataset.

| **Machine Learning Algorithm** | **Mean** | **SD** | **Best Parameters** |
|---|---|---|---|
| Logistic Regression | .867 | .001 | C= .01 |
| Random Forest Classifier | .892 | .011 | Max Depth = 1, Max Features = .5 |
| Support Vector Classifier | .892 | .012 | C= .001, Gamma= .001 |

After reviewing the output what I found most interesting was that the accuracy score did not change too much from the baseline. Additionally, the standard deviation from perturbation was 0 for all features. This reflected that these features had basically no impact to the baseline accuracy score. To confirm these findings, some side work was completed. First, I noticed the Kaggle Competition Users were also not getting score much above the baseline either. Second, the F-1 score was also close to 0 which again reflect there was basically no predictive power from the model or features.

## 5.0 - Outlook:

Although an in-depth analysis was completed to thoroughly approach the health insurance cross validation data set, there are limitations to this modeling approach. First, it was decided early on to ignore some parts of the dataset, so the dataset was more manageable because I was not participating in the Kaggle Contest. By using the given “training data”, which resulted in splitting the given training data into train, validation, and test, different random seeds could be used for splitting. For more optimal analysis it would be ideal to have leveraged all of the data provided.

Additionally, to improve the model approach, it would be a good idea to have tried beyond the three machine learning algorithms used in the model approach. By testing with other techniques such as XGBoost, QDA, or Naïve Bayes, the test set may have been able to achieve a higher accuracy. In the future, if more time allows, one should improve upon model performance by tuning additional hyperparameters which can provide additional confidence in the predictions. Lastly, looking more into feature selection by performing a more comprehensive EDA process or even using various mixtures of features to produce additional features could help to enhance accuracy and provide more confidence in the predictions or even find that some features had some impact to interest in purchasing vehicle insurance.

## 6.0 - References:

- Dataset Source: Kumar, Anmol. “Health Insurance Cross Sell Prediction.” Kaggle.Com, 11 Sept. 2020, www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction. Accessed 12 Oct. 2020.
- GitHub Repository: https://github.com/romacoffin/Health-Insurance-Cross-Sell-Prediction
- Source: Gupta, Shailaja. “Pros and Cons of Various Classification ML Algorithms.” Medium, 28 Feb. 2020, towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6
- Source: M, Hossin, and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations.” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, 31 Mar. 2015, pp. 01–11, pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf, 10.5121/ijdkp.2015.5201.