Roma Coffin
Data 1030 Midterm Project
GitHub Repository

Health Insurance Cross Sell Prediction

Introduction:

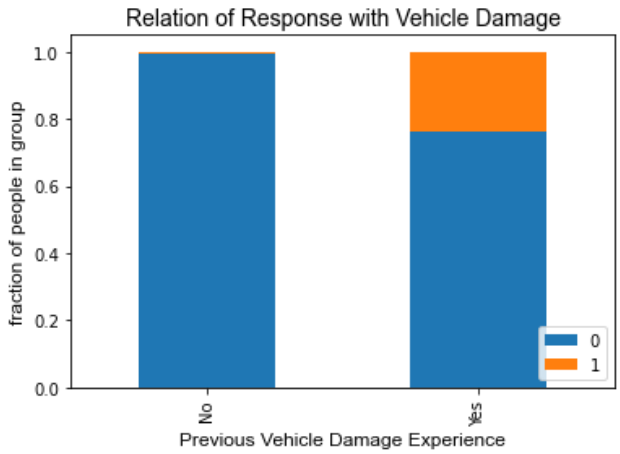
For my machine learning midterm project, I analyzed a data set that is aimed to see whether or not someone will be interested in purchasing vehicle insurance. The goal is for a health insurance provider to leverage their own customer's data from past years to see if they would also be inclined to buy 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, for each customer, this data will be split into train, validation, and test data. See the table below for a short description on each of the 12 features.

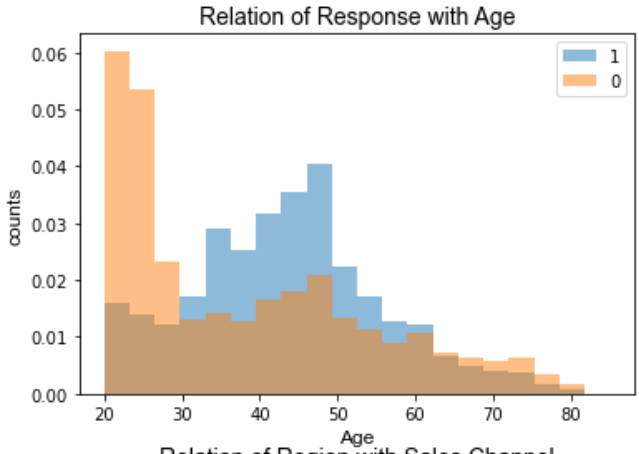
Feature	Numerical/Categorical	Description
Id	NA	Unique ID of each customer
Gender	Categorical	Male or Female
Age	Numerical	Age of customer
Driving License	Categorical-0,1	If they have a driver's license (0) or not (1)
Region Code	Numerical	Region customer is from (0-52)
Previously Insured	Categorical-0,1	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)

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.



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.



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.

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

Data preprocessing:

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. This is not time series data because it not indexed over a specific time period order. The data does not have a distinct group structure, so a basic split approach was performed. It is a large dataset but there are less than 1 million points, so the default allocation was used 60 train/20 val/20 test. There are now 10 features in the dataset, both Id and Vintage were removed because no impact to response variable was seen. See below on why preprocessor was chosen for each feature.

Feature	Numerical/Categorical	Preprocesser	Reason
Gender	Categorical	OneHotEncoder	Categorical can't be ordered
Age	Numerical	MinMaxEncoder	Numerical reasonably bounded
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

Reference:

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