# Health Insurance Cross Sell Prediction

October, 2021

Damilola Efuwape
Springboard

## Abstract

*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. 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 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?

## Specifications

For excellent analysis and prediction, the project will focus on demographic information like “Age”, “Gender”, etc, as well as vehicular information and policy types to determine which existing health insurance customers will be interested in auto insurance. This project will help stakeholders, management and decision makers of insurance companies track, analyse and assess the likelihood of getting potential vehicle insurance customers from the existing health insurance owners, thereby helping them plan their communication strategies to reach out to those customers, which will, in turn, optimize their business models and revenues. Machine learning models that will help make these predictions will be built. The project will look to provide answers to the following alternative questions amongst others:

1. Apart from rates, what other factors influence customers’ insurance providers decisions?

2. Why are there vast disparities among rates of different insurance companies?
3. What factors are used to determine rates for each customer?

## Data Collection

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 Processing, Preparation and Feature Engineering

The dataset does not have any missing values, but regular review will be done to make sure the data is in great shape. Data profiling and data mining will be done to get comprehensive information about the data and check for further defects. If there are any outliers, they will be capped with percentiles or replaced with the median values of their respective columns, depending on what percentages are missing.

## Machine Learning: Model Selection

XGBoost, RandomForestClassifier, Gaussian Mixture Clusters and some other models will be used for the project. Train-Test Split will be used to test the model performance and cross-validation will be used to test the model's performance.

When the project is complete, the results, along with a report and presentation slide will be uploaded to my GitHub repository.