**Vehicle Insurance Cross Sell Prediction**

**Problem Statement**

Our client is an Insurance company that has provided Health Insurance to its customers. Now they need help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

**Cross-selling**

Cross-selling is a sales technique involving the selling of an additional product or service to an existing customer. Cross-selling For example, if you encourage a customer who just bought a new phone to get a protective case at the same time, that's a cross-selling win.

Business Objective: An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalized in that year, the insurance provider company will bear the cost of hospitalization etc. for upto Rs. 200,000. Now if you are wondering how can a company bear such a high hospitalization cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes into picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalized that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year a customer needs to pay a premium of a certain amount to the insurance provider company so that in case of an unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue.

Now, in order to predict whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

Dataset: Dataset which contains a total of 381109 rows and 12 features. We have a categorical dependent variable Response which represents whether a **customer is interested in vehicle insurance or not**.

Features are Gender, driving license, region code, previously insured, vehicle age, vehicle damage, annual premium, policy sales communication channels, vintage: number of days customers associate with company. And **Target features is Response 2 class: 1 for Customer is interested, 0 for Customer is not interested.**

**Data Preprocessing**:

Check the Null values and duplicate in Data: no null and duplicate values present in our dataset

Feature Scaling: Normalize the numeric features, all features in same scale

**Exploratory Data Analysis**:

Numeric Features convert to Category features

Numeric Age feature, categorized **Age** as YoungAge, MiddleAge, and OldAge.

categorized the **Region\_Code** as Region\_A, Region\_B, Region\_C

categorized the **Policy\_Sales\_Channel** into channel\_A, channel\_B, channel\_C

And find observations with respect to target features

Customers belonging to youngAge are more interested in vehicle insurance.

Customers having vehicles older than 2 years are more likely to be interested in vehicle insurance

Customers having damaged vehicles are more likely to be interested in vehicle insurance.

**Feature Selection:**

For Feature Selection, we used Kendall's rank correlation coefficient for numerical features and for categorical features.: There is no correlation between these two features.

# **Feature Selection-How To Select Features Using Information Gain For Classification In ML: Mutual Information technique:** MI Estimate mutual information for a **discrete target variable**.

# Here we observed that Previously\_Insured is the most important feature and has the highest impact on the dependent feature

**Feature Transformation:**

**One-Hot Encoding**

One hot encoding is a process by which categorical variables are converted into numeric form that could be provided to ML algorithms to do a better job in prediction.

When there is not an ordinal relationship between variables, we use One-Hot Encoding.

With One-Hot Encoding the model doesn't assume a natural ordering between categories which may result in poor performance or unexpected results.

Conclusion:

Further, we applied Machine Learning Algorithms to determine whether a customer would be interested in Vehicle Insurance. For the Naive Bayes algorithm, we got an accuracy score of 68% and after hyperparameter tuning, the accuracy score increased to 72%. Similarly, for Decision Tree Classifier, AdaBoost, BaggingClassifier, LightGBM accuracy score was obtained around 82%-87%. So, we selected our best model as the model with an accuracy score of 85% considering precision and recall as we have an unequal number of observations in each class in our dataset, so accuracy alone can be misleading.