**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](https://en.wikipedia.org/wiki/Logistic_function) to model a [binary](https://en.wikipedia.org/wiki/Binary_variable) [dependent variable](https://en.wikipedia.org/wiki/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 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 is of the data is 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 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.