# Project Report

* The Auto Insurance data has 59 columns out of which one is id and one is target column.
* The data has 595212 rows.
* There are 57 distinct features.
* Data has 4 types of features
  + Binary
  + Categorical
  + Ordinal
  + Interval
* The target value is highly imbalanced. The minor class (1) has around 21694 rows whereas major class has (0) has 573518.
* There are few features which have very high number of missing values.
* For categorical and binary columns/features we used **mode** as missing value replacement for other we used **mean**.
* There are few columns which are highly correlated to each other, so removed one of them to be safe on curse of dimensionality.
* There were few outliers as well so removed those rows from data set so that we have better prediction.
* We have done standard scaling and one hot encoding (getdummies).
* There were 7 features with very low variance so removed those.
* Mainly used following models,
  + Logistic Regressor with imbalance and balanced
  + Randomforest Classifier with imbalance and balanced
  + LinearSVC
  + MLP Classifier with GridsearchCV