Analytics Olympiad 2022

Contest Link: https://machinehack.com/hackathons/analytics\_olympiad\_2022/

Problem Statement: The Vehicle Insurance business is a multi-billion dollar industry. Every year millions and millions of premiums are paid, and a huge amount of claims also pile up.

You have to step into the shoes of a data scientist who is building models to help an insurance company understand which claims are worth rejecting and the claims which should be accepted for reimbursement.

You are given a rich dataset consisting of thousands of rows of past records, which you can use to learn more about your customers’ behaviours. For example, you are supposed to create an ML model to look at a case of an insurance claim and decide whether to reject or accept it.

Columns: ['ID', 'AGE', 'GENDER', 'DRIVING\_EXPERIENCE', 'EDUCATION', 'INCOME',

'CREDIT\_SCORE', 'VEHICLE\_OWNERSHIP', 'VEHICLE\_YEAR', 'MARRIED',

'CHILDREN', 'POSTAL\_CODE', 'ANNUAL\_MILEAGE', 'SPEEDING\_VIOLATIONS',

'DUIS', 'PAST\_ACCIDENTS', 'OUTCOME', 'TYPE\_OF\_VEHICLE']

Learn and predict the OUTCOME variable.

Categorical features : 'AGE',

'GENDER',

'DRIVING\_EXPERIENCE',

'EDUCATION',

'INCOME',

'VEHICLE\_OWNERSHIP',

'VEHICLE\_YEAR',

'MARRIED',

'CHILDREN',

'TYPE\_OF\_VEHICLE' are converted to numerical features using one-hot encoding technique.

Note: POSTAL\_CODE is an another 5 digit numerical feature, which to be treated a categorical feature , and have taken each individual digit of 5 digits as a feature, and then one hot encoded each of these 5 features(As of each them would contain 0-9 as value). This would enable us to capture latitude and longitude information.

Class imbalance is handled using stratified k fold cross validation.