This document will give an overview of the main steps involved in our data science project.

We first performed the preprocessing and cleaning of our dataset before proceeding to the exploratory data analysis of our data to derive insights. Finally, using the cleaned data we built our Machine learning model to predict claims or not on the test set.

The major steps from each phase will be highlighted below.

1 Preliminary processing of data

This phase deals with cleaning the raw data and making it ready for analysis and Machine learning modeling.

In this section we have performed the following tasks sequentially:

1. Perform necessary package installations (on top of anaconda default).
2. Read our training dataset into a pandas dataframe.
3. Initial look at our dataset.
4. Remove rows with empty elements.
5. Remove duplicate rows.
6. Write the cleaned dataframe to a new CSV file (**train_cleaned.csv**).

The code with each section commented appropriately is contained in the notebook titled: **'Preliminary-processing.ipynb'**.

2 Exploratory data analysis (EDA)

This phase deals with exploring our cleaned data (**train_cleaned.csv**) in order to derive useful insights from our data and if needed gauge the importance of each variable with respect to the target and see how each independant variables are related to each other.

In this section we have performed the following tasks sequentially:

1. First find the distribution of each of the variable - categorical and numerical. Deduce appropriate inferences.
2. Find the influence of each variable on the target variable - numerical and categorical. Plot the most important ones against the target to visualise. Draw appropriate inferences.
3. Relationships between the independant variables.
   (a) Numerical attributes amongst themselves.
      i. Correlation is measured between the numerical variable pairs.
      ii. The correlated pairs are plotted against each other for closer inspection.
      iii. Deduce inferences from the visualisations.
   (b) Categorical attributes amongst themselves.
      i. Correlation concept does not exist for nominal variables.
      ii. Pairs are plotted against each other to visualise possible relations between them.
      iii. Deduce inferences from the visualisations.
   (c) Analysis of relationship between categorical and numerical attributes.
      i. Plot certain categorical variables of interest against numerical attributes.
      ii. Deduce inferences from the visualisations.
4. Detail the next steps.

The code with each section commented appropriately is contained in the notebook titled: **'EDA.ipynb'**.

3 Machine learning modeling

Now that we have processed our data and explored it through different visualisations and statistical metrics, for this phase we will use the cleaned dataset to fit machine learning models on our training data.

In this section broadly we have performed the following tasks sequentially:

1. Preparing the data for training the machine learning models.
2. Fitting machine learning models on the training data.
   (a) Intrinsic feature selection models.
      i. Types of Intrinsic feature selection models.
      ii. Parametric modeling (example: Logistic Regression with L1-penalty) .
         A. Cross-validation on Logistic regression to measure model performance.
         B. Fitting the final model.
         C. Predict the probabilities of the class prediction and plot the ROC curve.
      iii. Non-Parametric modeling (example: XGBoost).
         A. Cross-validation on XGBoost to measure model performance.
         B. Fitting the final model.
         C. Predict the probabilities of the class prediction and plot the ROC curve.
      iv. Comparision of results between Logistic Regression and XGBoost.
      v. Note on improving performance.
   (b) Manual feature selection based on performance result (Wrapper method).
   (c) Manual feature selection based on metric result (Filter method).
   (d) Dimensionality Reduction for Feature extraction then modeling.
   (e) Improving the modeling.
3. Applying the learned models to the test data.
   (a) Extract the test data and clean it.
      i. Read the data to a dataframe
      ii. Drop the target columns which are empty.
      iii. Index stored seperately.
      iv. Removing the dollar symbol in the anomalous colomns.
      v. Arrange all the categorical colomns and numerical colomns together.
      vi. Remove all the empty rows.
      vii. Remove duplicate rows if any.
   (b) Encoding the categorical variables.
      i. Perform one-hot encoding.
      ii. Verify column order match for the prepared training and test data.
   (c) Logistic regression predictions on the processed test data.
      i. Prediction of class output.
      ii. Prediction of probabilities assosciated for each class prediction.
   (d) XG-Boost predictions on the processed test data.
      i. Prediction of class output.
      ii. Prediction of probabilities assosciated for each class prediction.
   (e) Note on the results of the modeling process.
4. Suggested the next steps/improvements.

The code with each section commented appropriately is contained in the notebook titled:
'Machine_Learning_Modeling.ipynb'.
In this notebook we have mentioned the loss and score metrics for both the models - Cross-Entropy for loss and Accuracy, Precision, Recall and F1-score along with the ROC-AUC score for evaluation of performance (scoring).

The output class predictions are stored in **'Logistic_Regression_test-classifications.csv'** for the logistic regression model and **'XGBoost_test-classifications.csv'** for the XGBoost model.

We have also calculated the probabilities for each class prediction for both the models and written them to a file. **'Logistic_Regression_test-probabilities.csv'** for the logistic regression and **'XGBoost_test-probabilities.csv'** for the XGBoost model.