Welcome to Team 16’s video about the insurance dataset and our understanding and throught process regarding this dataset provided.

Automotive insurance has a lot of factors involved, from a persons age to the area a person lives in is all detrimental to pricing a policy, but the biggest is predicting whether a policyholder will ever get into an accident and incur a loss to the insurance company?

In the dataset provided to us we have over 60 columns and 420,000 rows. Not all of the 60 columns provided to us actually make an impact on the loss amount predicted, to select what columns to select we first one hot encode the categorical data which further increases column size and then apply feature correlation and pick the 30 highest correlated features with respect to Loss\_Amount.

We check for outliers in non categorical variables if any and use nearest neighbors method with similar features and scale down the outlier. After that we perform a sort of compression of the data, we take 500 rows and summarise them with averages and min-max’s to reduce the number of rows we have, this reduces the number of rows to around 820, after performing feature scaling we can move on to modelling the data for predictions.

With our clean and compressed data on the training set, we proceed to do the same for the testing set for each portfolio, reducing the portfolio entries to either 1-6 columns, for training we used a set of different algorithms varying from each other in nature, algorithms such as Ridge, Lasso, XGBoost, Light Gradient Boosting, Gradient Boosting, Kernel Ridge and Elastic Net. We take predictions from all the trained models and get the average loss of each portfolio, then calculate the Loss Ratio and take its natural logarithm.