**Capstone Project 1: In-Depth Analysis**

**(House Price Prediction)**

To predict the house price of the given set of data, I used supervised machine learning techniques. The aim of the model is to obtain the minimum RMSE (Root Mean Square Error) on the test data. I build multiple regression models to predict the house price and calculated RMSE for each model on the Test Data. I tested the model on the test data in Kaggle and received a RMSE 0.1301.

Terminologies used:

Given data with dependent variable is divided as train set and test set and data without dependent variable is called test data.

First, I combined the test data and train data to perform Data Pre-Processing and Exploratory analysis. Once the data is clean and processed, the data is split into train data and test data. Then the train data is divided into independent as x and dependent as y. There are different methods to measure the performance of the regression models such as Mean Squared Error, Root Mean Squared Error, R-Squared score and Mean absolute deviation. I used RMSE (Root Mean Square Error) as the metric for measuring the performance of the model. I normalized the train and test data using robust scaler transformation. I again split the training data into 80:20 as training and testing set. This model learns on the 80% of the train data and tests on the 20% of the unseen data before finally testing on the real test data.

Secondly, I build the Linear Regression as the base model with default parameters, and I fit the train set to the model and predicted the house price on the test set and calculated metrics such as RMSE, R2\_score to check the performance of the model. The RMSE when predicted house price on the test set is 0.1545 and when submitted in the kaggle website on the test data with this model is 0.224.The coefficients of the model can be obtained using” .coef\_ “and intercept using” .intercept\_”. Then I used model Ridge regularization on the dataset with hyperparameter optimization. Using cross validation techniques such as GridSearchCV and RandomSearchCV to tune the hyperparameters to improve the performance of the model.Then the same metrics are used on the Ridge regression and obtained RMSE of 0.102 on the test set and 0.15.

Similarly the data is fitted into multiple models such as BayesianRidge, KernelRidge, Lasso, ElasticNet, KNNRegressor, DecisionTreeRegressor, SVMRegressor and Boosting algorithms .The metrics are calculated and submitted in Kaggle. Bayesian Ridge gives the best RMSE of 0.130 among all the models. Still more hyper parameter tuning to be done for better accuracy on the data.