**REGRESSION**

1. Regression is the process of predicting a continuous value. There is the Dependent variable which is usually denoted by Y which derived from the independent Variables usually denoted by X. There is linear regression and nonlinear regression.
2. Linear Regression is a type of regression in which there is an approximation of a linear model used to describe the relation between two or more values. For example, for a data set df which has historical data of Engine size Cylinders and CO2 emissions of different models of cars we can predict the CO2 emissions of a car with the use of the Engine size or and the Cylinders. This method is very fast as it involves no parameter tuning

Linear regression is only applicable when the dependent variable is a continuous value and not binary

1. Metrics evaluation of Regression

Mean Average Error (MAE)

Mean Squared Error (MSE)

1. Model bias is how best suited a model is to a training set

Model variance is how the model changes when you use a different part or section of the dataset to train it

Bias variance tradeoff is when actions that are taken to decrease the bias of the model inherently increases the variance and vice versa.

**VALIDATION**

1. Over-fitting is a situation where the model is overly trained to the dataset which may cause noise and as a result a model which is not generalized
2. We validate the model using basically two approaches
3. Tran and Test on the same dataset: In this approach we use the same dataset to test the model after using it to train the model. This may sometimes result in over-fitting because there is a likelihood of the model producing high training accuracy
4. Train/test Split: Here, we split into parts, the dataset into training and test data. This improves out of sample accuracy
5. We need to split data to improve our out of sample accuracy. This is the number of correct predictions the model can make when it’s being tested with data it has not been trained with.

I will split Data into two namely: Train Data and Test Data.

1. Cross Validation is an approach which seeks to reduce over fitting in a model. Works by making a fixed number of folds of the dataset and running analysis on each fold and averaging your findings. For example, if we divide a dataset into 2 folds then we test the model with 1st Fold of the dataset and we train with remaining after we repeat same process with next half. Accuracy of both Folds is recorded and an average is calculated

If we have k= 4folds then we split dataset into 4. First quarter of the dataset is used to test the model and the remaining is used to train the model. The accuracy is calculated and recorded, in the second fold the next quarter is used to test the data and the remaining used for training. The accuracy is also recorded. We repeat this for all the quarter and the accuracy is averaged to produce a more consistent result.

**CLASSIFICATION**

1. Classification is a supervised learning approach where there is categorization of some unknown items into discrete set of categories or classes.

K-Nearest Neighbor

Decision Trees

Support Vector Machines

1. Logistic Regression is a classification algorithm for categorical variables. It is more applicable when the data is binary. Like in a dataset df where we use its historical data to predict whether or not a customer will not renew subscription to a particular service. If for instance we are to use to use other variables to predict the amount customers are likely to spend on a new product which is a continuous value then logistic regression will not be effective.
2. Logistic Regression is a linear model because the outcome is the sum of the model’s inputs and parameters
3. We evaluate classifications by using one of the following Classification Evaluation metrics
4. F1-score
5. Jaccard index
6. Accuracy as its name suggest is how many predictions the model is able to correctly predict.
7. A confusion table is a set of matrix which displays counts of True Positives, True Negatives, False Positives and False Negatives. Cells are made up of True Positives, True Negatives, False Positives and False Negative
8. Precision evaluates how precise a model is.
9. Recall is how many true positives a model correctly predicts.
10. F1-score is a measure we use when we need to create a balance between recall and precision.