Machine Learning Techniques/Algorithms :

1. Regression
2. Classification
3. Clustering

Linear Regression :

1. Simple linear regression
2. Multiple Linear Regression

ML Model:

Basic of simple linear regression:

Find the best fitted line for a model and all the various parameters associated with it.

Introduction to Machine learning

Supervised and unsupervised Methods

The linear regression model

Residuals

Residual sum of square(RSS) and R2 (R-Squared)

CRISP DM- Framework

Regression: Output variable to be predicted is a **Continuous Variable**

Classification: Output variable to be predicted is a **Categorical Variable.**

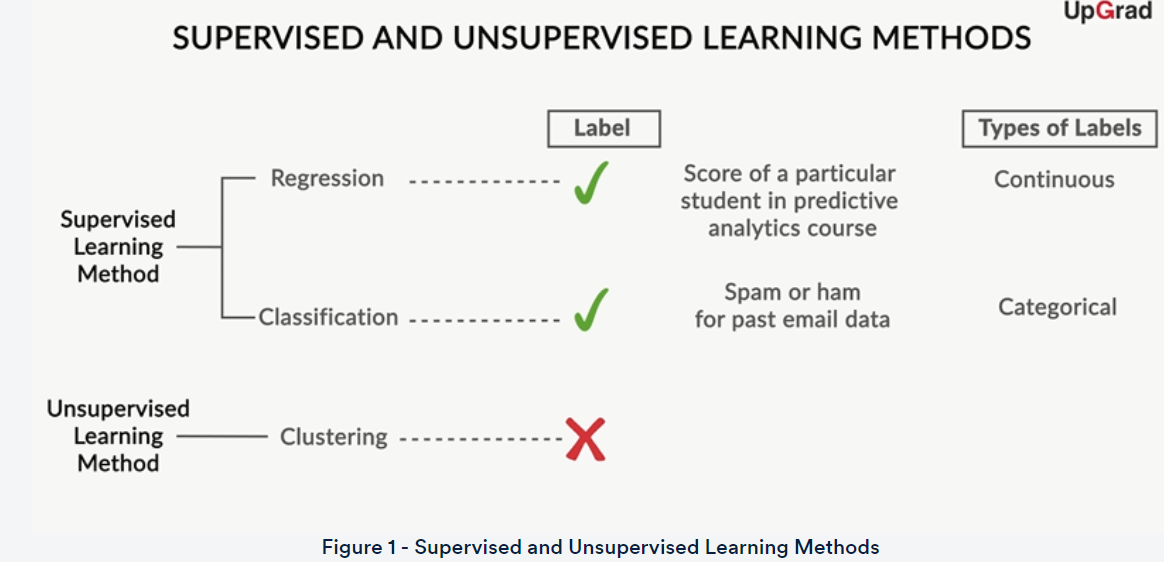
Clustering: No predefined notion of label is allocated to the groups/clusters formed

Predicting the score of a student (example for Regression)

Predicting the email as spam or Ham (Example of Classification)

Customer segmentation(example of clustering)

Predict the label using the model built on past data : Model built with Past data is used for predicting the label for upcoming or test data or new data.



Ei = Yi-Ypred

B0+b1(xi) – (Bo+b1(xpred)

Ei = B1(xi-xpred)

Summary Of Basic Machine Learning:

Machine learning models can be categorized based on the learning algorithm as supervised and unsupervised learning. Past data with labels is available to build a model in the supervised learning whereas past data with labels is not available so model has to segment the labels into clusters or groups as there are no predefined labels.

Supervised learning methods are categorised based on the output label type continuous and categorical as Regression and Classification algorithm.

A Dataset is divided into training data and testing data in the supervised learning method. Training data is used for the model to learn whereas testing data is used for predicting or testing the model.

Linear regress

1. Machine learning models can be classified into the following two categories on the basis of the learning algorithm:
   * **Supervised learning method:**Past data with labels is available to build the model.
     + **Regression:**The output variable is continuous in nature.
     + **Classification:**The output variable is categorical in nature.
   * **Unsupervised learning method:**Past data with labels is not available.
     + **Clustering:** There is no predefined notion of labels.
2. Past dataset is divided into two parts in the supervised learning method:
   * **Training data**is used for the model to learn during modelling.
   * **Testing data**is used by the trained model for prediction and model evaluation.
3. Linear regression models can be classified into two types depending upon the number of independent variables:
   * **Simple linear regression:** This is used when the number of independent variables is 1.
   * **Multiple linear regression:** This is used when the number of independent variables is more than 1.
4. The equation of the best fit regression line Y = β₀ + β₁X can be found by minimising the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:
   * **Differentiation**
   * **Gradient descent**
5. The strength of a linear regression model is mainly explained by R², whereR² = 1 - (RSS/TSS).
   * **RSS:** Residual sum of squares
   * **TSS:** Total sum of squares
6. RSE helps in measuring the lack of fit of a model on a given data. The closeness of the estimated regression coefficients to the true ones can be estimated using RSE. It is related to RSS by the formula: RSE=√RSSdf, where df=n−2 and n is the number of data points.

**Simple Linear Regression:**

1. Assumptions of simple linear regression
   * Linear relationship between X and y.
   * Normal distribution of error terms.
   * Independence of error terms.
   * Constant variance of error terms.
2. Hypothesis testing in linear regression
   * To determine the significance of beta coefficients.
   * H0:β1=0;HA:β1≠0.
   * T-test on the beta coefficient.
   * t score=^βiSE(^βi).
3. Building a linear model
   * OLS (Ordinary Least Squares) method in statsmodels to fit a line.
   * Summary statistics
     + F-statistic, R-squared, coefficients and their p-values.
4. Residual Analysis
   * Histogram of the error terms to check normality.
   * Plot of the error terms with X or y to check independence.
5. Predictions
   * Making predictions on the test set using the 'predict()' function.
6. Linear Regression using SKLearn
   * A second package apart from statsmodels for linear regression.
   * A more hassle-free package to just fit a line without any inferences.

Simple Linear Regression:

Visualize the relation between X and y variable using scatter plots and observe for linearity.

Validate the assumptions of linearity(linear relationship between X and y,normality of error terms, independence of error terms, constant variance of error terms)

Find thesignificance using Hypothesis testing for significance of B coefficients

Fit the line using stats models and find the significance using F-test

Residual analysis and predictions

two basic ways of dealing with multicollinearity:

1. Looking at **pairwise correlations**
   * Looking at the correlation between different pairs of independent variables
2. Checking the **variance inflation factor** (VIF)
   * Sometimes, pairwise correlations are not enough.
   * Instead of just one variable, the independent variable may depend upon a combination of other variables.
   * VIF calculates how well one independent variable is explained by all the other independent variables combined.

The VIF is given by:

                                                                               VIFi=1/1−Ri2

Here,*'i'* refers to the i-th variable, which is being represented as a linear combination of the rest of the independent variables. You will see VIF in action during the Python demonstration on multiple linear regression.

The common heuristic we follow for the VIF values is:

**> 10:** VIF value is definitely high, and the variable should be eliminated.

**> 5:** Can be okay, but it is worth inspecting.

**< 5:**Good VIF value. No need to eliminate this variable.

Some methods that can be used to deal with multicollinearity are as follows:

1. **Dropping variables**
   * Drop the variable that is highly correlated with others.
   * Pick the business interpretable variable.
2. **Creating a new variable** using the interactions of the older variables
   * Add interaction features, i.e., features derived using some of the original features.
3. **Variable transformations**
   * Principal component analysis (covered in a later module)

**Model Assessment and Comparison**

Once the model is built, you would want to assess it in terms of its predictive powers. For multiple linear regression, you may build more than one model with different combinations of the independent variables. In such a case, you would also need to compare these models with one another to check which one yields optimal results.

Besides, selecting the best model to obtain decent predictions is quite subjective. You need to maintain a balance between **keeping the model simple** and **explaining the highest variance** (which means that you would want to keep as many variables as possible). You can do this using the key idea that a model can be penalised for keeping a large number of predictor variables.

Hence, there are two new parameters that come into the picture:

                                                              Adjusted R2=1−(1−R2)(N−1)N−p−1

                                                                  AIC=n×log(RSSn)+2p

Here, n is the sample size, meaning the number of rows you would have in the data set, and p is the number of predictor variables.

**R-Squared vs Adjusted R-Squared**

**Why, according to you, is it better to use adjusted R-squared in multiple linear regression?**

The major difference between R-squared and adjusted R-squared is that R-squared does not penalise the model for having more number of variables. Thus, if you keep on adding variables to the model, the R-squared will always increase (or remain the same when the value of the correlation between that variable and the dependent variable is 0). Thus, R-squared assumes that any variable added to the model will increase the predictive power.

Adjusted R-squared, on the other hand, penalises models on the basis of the number of variables present in them. So, if you add a variable and the adjusted R-squared drops, you can be certain that that variable is insignificant to the model and should not be used. Thus, in the case of multiple linear regression, you should always look at the adjusted R-squared value in order to keep redundant variables out of your regression model.

Feature Selection

The one crucial aspect of multiple linear regression that remains to be discussed is feature selection. When building a multiple linear regression model, you may have quite a few potential predictor variables; selecting just the right ones becomes an extremely important exercise.

Let’s see how you can select the optimal features for building a good model.

Note that in the brute force approach, one of the combinations out of 2pdoes not make any sense, namely the one which does not use any features at all.  So, this means that we only need to try 2p - 1 feature combinations.

To get the optimal model, you can always try all the possible combinations of independent variables and see which model fits best. But this method is time-consuming and infeasible. Hence, you need another method to get a decent model. This is where manual feature elimination comes in, wherein you:

1. Build the model with all the features,
2. Drop the features that are the least helpful in prediction (high p-value),
3. Drop the features that are redundant (using correlations and VIF),
4. Rebuild the model and repeat.

Note that the second and third steps go hand in hand, and the choice of which features to eliminate first is very subjective.

Now, manual feature elimination may work when you have a relatively low number of potential predictor variables, say, ten or even twenty. But it is not a practical approach when you have a large number of features, say 100. In such a case, you automate the feature selection (or elimination) process. Let's see how.

* RFE

brief summary of what you learnt in this session:

1. When one variable might not be enough
   * A lot of variance values are not explained by just one feature.
   * Predictions are inaccurate.
2. Formulation of MLR
3. New considerations to be made when moving from SLR to MLR
   * Overfitting
   * Multicollinearity
   * Feature selection
4. Dealing with categorical variables
   * Dummy variables for fewer levels
5. Feature scaling
   * Standardisation
   * MinMax scaling
   * Scaling for categorical variables
6. Model assessment and comparison
   * Adjusted R-squared
   * AIC, BIC
7. Feature selection
   * Manual feature selection
   * Automated feature selection
   * Finding a balance between the two

Top3 takeaways from this session:

* For model comparison and evaluation we use adjusted r2 as a measure which penalises for using more number of features, keep feature which explains the maximum variance and at the same time it should be simple and not overfitting.
* To drop a feature we can look at the p value, study the VIF and correlations
* For feature selection we can use manual feature elimination when number of features are low but when they are high we go for automated approach called, RFE or Based on AIC

**Multiple Linear Regression Summary:**

**Summary**

Let’s summarise what you learnt:

1. It is important to understand if a linear regression modelling will be applicable to the problem you are trying to solve. For example, linear regression cannot help you decide if a customer will opt out of a subscription, as this is a classification problem.
2. Linear regression guarantees interpolation but not extrapolation.
3. While linear regression can be used for both projection and prediction, there is a difference between the two. In prediction, the goal is to identify the most important variables that explain the outcome in a simpler way. In projection, the goal is to accurately forecast the outcome, no matter how complex the model gets.
4. The business goal is crucial and will decide what path the modelling process should take.
5. In the industry, variables that are actionable are valued over others. For example, given two quite similar variables, “Views to the platform” and “Visitors to the platform”, the latter is more actionable, as it is easier to get viewers to the platform than forcing anybody to watch the shows.
6. You don’t end the modelling process until you are sure that no more significant variables could be added to explain the outcome. Thus, you check for randomness of errors, which could indicate whether any KPI that could have helped explain the outcome was left out.