Questions week4

1/Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. What algorithms can be used to automatically select the most important features (regression, etc..)? Describe at least 3?

Random forest provides the branches in the order of the most important branch first

P scores on the lsm models show which parameter has the most weight

Lasso or ridge regressions which are using regularization to decrease the coefficient which are not important

2/Explain data leakage and overfitting (define each)?  
Explain the effect of data leakage and overfitting on the performance of an ML model.

Data leakage occurs when model is trained on data which are not on the from a specific time and tested on data from another time.

Overfitting is when your training model has been built to fit all the points from the training perfectly. It did learn the pattern of the residuals. Applied to test data – or other data - it performs poorly because the residuals will be different.

3/Explain what are our outliers in your data?  
Explain at least two methods to deal/treat outliers in your data?

Outliers are data which are very different from the other data. They could be very large, or very small and are not very many. It can be identified with visualization (box plot) / calculating the distance from the mean. It can then be removed or replace by some average that can represent them (EX all values registered on a Wednesday=0, take the average values of the previous Wednesday)

4/What is feature scaling and why is it important to our model?  
Explain the different between Normalization and Standardization?

Feature scaling is a way to give each feature the same importance as they are all in the same scale

Normalization is to set values between 0 and 1

Standardization is to make the data to have a mean of 0 and a standard deviation of 1.