Machine Learning Assignment 4

A1) option C- between -1 and 1

A2) option C- Recursive feature elimination

A3) option A- Linear

A4) option A-Logistic Regression

A5) option B- same as old coefficient of 'X'

A6) option B- increases

A7) option C- Random Forests are easy to interpret

A8) option B & C

A9) option B,C & D

A10) option A,B & D

A11) Outlier- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Inter Quantile Range (IQR)- IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

A12) Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data.

Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification. It attempts to increase the weight of an observation if it was erroneously categorized. Boosting creates good predictive models in general.

A13) Adjusted R^2 indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

A14)

Normalization	Standardisation
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers
Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

A15) Cross Validation- Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

Advantage- In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage- Cross Validation drastically increases the training time. Earlier we had to train our model only on one training set, but with Cross Validation we have to train our model on multiple training sets.

SQL Worksheet 4

A1) SELECT AVG (orderShipped), shippedDate FROM orders WHERE orderShipped=(SELECT COUNT(shippedDate) FROM orders GROUP BY shippedDate);

A2) SELECT AVG (orders), orderDate FROM orders WHERE orders=(SELECT COUNT(orderNumber) FROM orders GROUP BY orderDate);

A3) SELECT productName, pro_price FROM products WHERE pro_price = (SELECT MIN(MSRP) FROM products);

A4) SELECT productName, proQuantity FROM products WHERE proQuantity = (SELECT MAX(quantityInStock) FROM products);

A5) SELECT p.productName FROM orderdetails AS o INNER JOIN products AS p ON o.productCode= p.productCode GROUP BY o.productCode
ORDER BY SUM(o.quantityOrdered) DESC;

A6) SELECT c.customerName FROM payments AS p INNER JOIN Customers AS c ON p.customerNumber= p.customerNumber GROUP BY p. customerNumber ORDER BY SUM(p.amount) DESC LIMIT3;

A7) SELECT customerNumber, customerName FROM Customers WHERE city='Melbourne';

A8) SELECT customerName FROM Customers WHERE customerName LIKE 'N%';

A9) SELECT customerName FROM Customers WHERE city='LasVegas' AND phone LIKE '7%';

A10) SELECT customerName FROM Customers WHERE creditLimit<10000 AND city='Las Vegas' OR city= 'Nantes' OR city='Stavern';

A11) SELECT orderNumber FROM orderdetails WHERE quantityOrdered<10;

A12) SELECT orderNumber FROM orders WHERE customerName LIKE 'N%';

A13) SELECT customerName FROM orders WHERE status=’Disputed’;

A14) SELECT customerName FROM payments WHERE checkNumber LIKE ‘H%’ AND paymentDate=’ 2004-10-19’;

A15) SELECT checkNumber FROM payments WHERE amount>1000;

Statistics Worksheet 4

A1) Central Limit Theorem and its importance-

The CLT is a statistical theory that states that if we take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution.

A2) Sampling- Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed.

Methods of Sampling- There are usually two methods of sampling, they are:

1) Probability Sampling-It involves random selection, allowing us to make strong statistical inferences about the whole group.
2) Non-probability sampling- It involves non-random selection based on convenience or other criteria, allowing us to easily collect data.

A3) A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

A4) A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

If such a data set is plotted then the graph looks like a bell shaped curve.

A5) Covariance- It is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a

change in one variable. The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship. The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number. A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables.

Correlation- It is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated.

A6) Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

A7) Sensitivity Analysis- It is a financial model that determines how target variables are affected based on changes in other variables known as input variables. It is a way to predict the outcome of a decision given a certain range of variables.

We can determine sensitivity by following these steps:

1. Firstly the base case output is defined.
2. Then the value of the output at a new value of the input while keeping other inputs constant is calculated.
3. Find the percentage change in the output and the percentage change in the input.
4. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

A8) Hypothesis Testing- Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). One of these is the claim to be tested and based on the sampling results which infers a similar measurement in the population, the claim will either be supported or not. The claim might be that the population proportion (or mean) has increased, decreased, stayed the same, or that it has changed. According to the words used in the problem, the claim will be either H0 or H1.

A9) Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables. For example: Age, Height, Weight etc.

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. For example: Gender, Religion, Marital status etc.

A10) Range- The range is the easiest measure of variability to calculate. To find the range, we need to follow these steps:

1. Order all values in your data set from low to high.
2. Subtract the lowest value from the highest value.

The mathematical formula for calculating range is:

$$R=H-L$$

- R = range
- H = highest value
- L = lowest value

Inter Quantile Range (IQR) - The Interquartile Range (IQR) formula is a measure of the middle 50% of a data set. The smallest of all the measures of dispersion in statistics is called the Interquartile Range. The difference between the upper and lower quartile is known as the interquartile range.

The formula for calculating IQR is:

Interquartile range = Upper Quartile – Lower Quartile

$$IQR=Q3-Q1$$

were,

IQR = Interquartile range

Q1 = (1/4)[(n + 1)]th term)

Q3 = (3/4)[(n + 1)]th term)

n = number of data points

A11) A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.

A12) We usually use Statistical tests (z scores) for identifying outliers.

A13) The *p*-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. *P*-values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the *p*-value, the more likely you are to reject the null hypothesis.

A14) Binomial probability function refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes. If the probability of success on an individual trial is p , then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$.

A15) ANOVA- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables. If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.