ASSIGNMENT - 3

2451-18-733-001
M. Ramani Priya

1. Define Predictive modeling. List out the types of Predictive modeling

→ Predictive modeling is a statistical technique using machine learning & data mining to predict and forecast likely future outcomes with the aid of historical & existing data.

→ A predictive model is not fixed, it is validated & revised regularly to incorporate changes in the underlying data.

Types of Predictive models.

1. Classification model:
   It categorises data for simple & direct query response

2. Clustering model:
   This model categorises data together by common attributes. It works by grouping things or people with shared characteristics or behaviours & plans strategies for each group with at a larger scale.

3. Forecast model:
   This model works on anything with a numerical value based on learning from historical data.

4. Outliers model:
   This model works by analyzing abnormal or anything or outlying data points.

5. Time series model:
   This model evaluates a sequence of data points based on time.

Common Predictive algorithms

1. Random Forest
2. Generalized Linear Model (GLM) for 2 values.
3. Gradient Boosted model
4. K-means.
5. Prophet.

2. Write the possible ways of improving accuracy of Linear regression.

Possible ways of improving accuracy of Linear regression.

1. Add more data

It allows "data to tell for itself", instead of relying on assumptions & weak correlations.

2. Treating missing & outlier values.

Unwanted presence of missing & outlier values leads to in training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. So it is important to treat missing & outlier values well.

Dealing with missing & outlier values:

1. Missing: In case of continuous variables, you can impute the missing values with mean, median & mode.

For categorical variables, the variables can be treated as seperate class.

2. Outlier: The observations can be deleted, perf transformed, binning, Imputation or treat outliers seperately.

3. Feature engineering

It helps to extract more information from existing data.
New information can be extracted interms of new features.
These features may have higher ability to explain the variance in training data.
Feature engineering is highly influenced by hypothesis generation.
It can be divided into 2 steps:

1. Feature transformation.
2. Feature Creation.

4. Feature Selection.

is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target variable.
The features can be selected based on various metrics like,

1. Domain knowledge
2. Visualization
3. Statistical parameters.

5. Multiple algorithms

Hitting at right machine learning algorithm is the ideal approach to achieve higher accuracy.
Some algorithms are better suited to a particular set of data sets than others.

6. Algorithm Tuning

The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model.

To tune these parameters you must have a good understanding of these meaning & their individual impact on model.

7. Ensemble the methods

This technique simply combines the result of multiple weak models & produce better results. This can be achieved through many ways:

1. Bagging
2. Boosting

8. Cross validation:

Cross validation is one of the most important concepts in data modeling.

This ~~Model~~ method helps to achieve more generalised relationships.

2451-18-733-001

3. What is Linear regression? List out critical assumptions of Linear regression?

→ Linear regression models the expected value of a numeric quantity (called the dependant variable) in terms of numeric and categorical inputs (called independant variable) or explanatory variables

In general terms, y[i] is the numeric quantity we want to predict & x[i,] is a row of inputs that corresponds to y[i].

Linear regression finds a fit function f(x) such that

$$y[i] = f(x[i,]) = b[1]\,x[i,1] + \cdots + b[n]\,x[i,n]$$

b[1], b[2], ... b[n] (called coefficients/beta) such that f(x[i,]) is as near as possible to y[i]

In an idealised theoritic situation,

$$y[i] = b[1]\,x[i,1] + b[2]\,x[i,2] + \cdots + b[n]\,x[i,n] + e[i]$$

this means, y is linear ~~corr~~ in the values of x i.e., a change in value of x[i,m] by one unit (while holding all other x[i,k]s constant) will change the value of y[i] by the amount b[m] always, no matter what the starting value of x[i,m] was.

The last term, e[i] represents unsymet unsystematic errors or noise. Unsystematic errors average to 0 & are uncorrelated with x[i] & y[i].

There are 4 assumptions associated with Linear regression model

1. Linearity: The relationship between X & the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.

3. Independence : Observations are independent of each other.

4. Normality : For any fixed value of X, Y is normally distributed

4. Write in detail about Linear and Logistic regression.

Linear regression model.

- Linear regression models the expected value of a numeric quantity (called the dependent or response variable) in terms of numeric & categorical inputs (called the independent or explanatory variable).

In general, suppose that y[i] is the numerical quantity we want to predict, & x[i,] is a row of inputs that corresponds to output y[i].

Linear regression finds a fit function f(x) such that,

$y[i] = f(x[i,]) = b[1] x[i,1] + \cdots + b[n] x[i,n]$

b[1], b[2], ... b[n] - coefficients or betas.

f(x[i,]) is as near as possible to y[i] $\forall$ (x[i,], y[i]) pairs

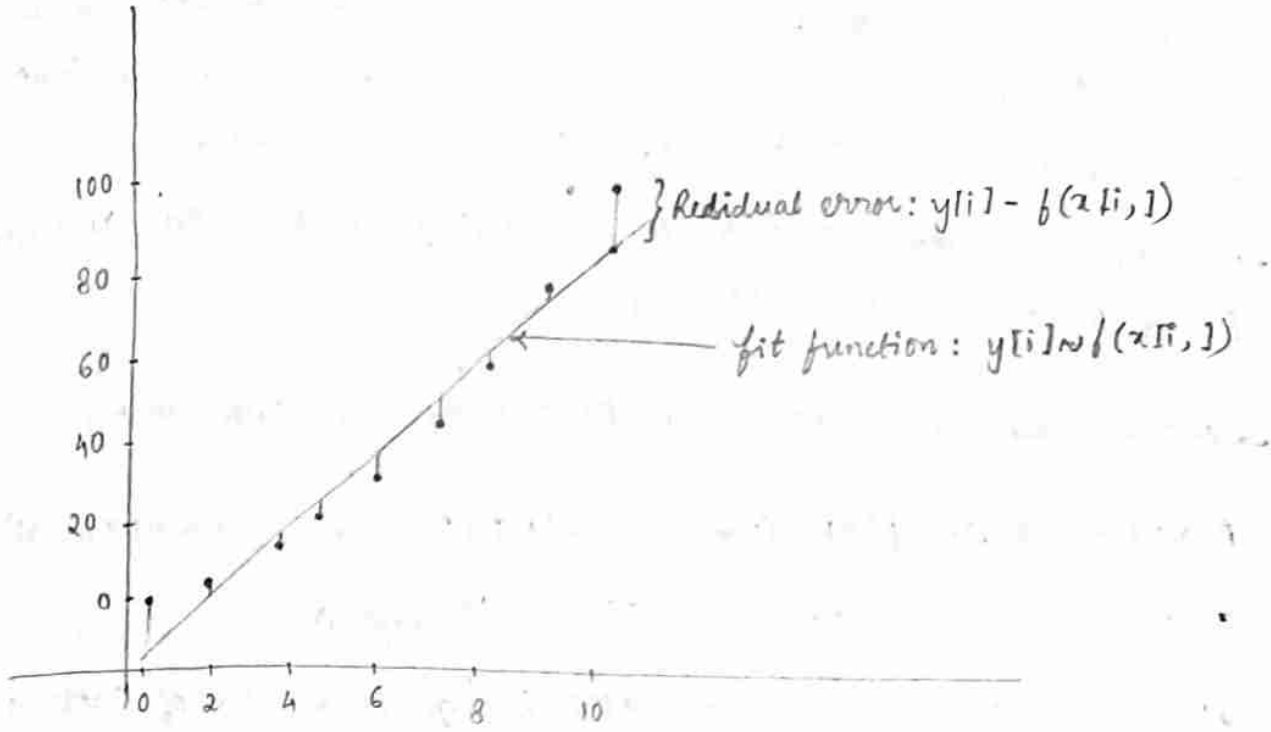
In an idealized theoretic situation,

$y[i] = b[1] x[i,1] + b[2] x[i,2] + \cdots b[n] x[i,n] + e[i]$

e[i] - unsystematic errors or noise.

for example.

$y = x^2$



Fit versus actual for $y = x^2$

→ All of the information in a linear regression model is stored in a block of numbers called coefficients.

→ In linear regression, the coefficients are chosen to minimize the sum of squares of the residuals. The method used is called the least squares method.

→ Interpreting model significances:
Most of the tests of linear regression, including the tests for coefficient & model significance are based on the error terms or residuals are normally distributed.
It's important to examine graphically ~~if~~ or using quantile analysis to determine if regression model is appropriate.

Logistic regression:

It is a member of a class of models called generalised linear models. Logistic regression can directly predict values that are restricted to the (0,1) interval, such as probabilities.

→ Logistic regression predicts the probability y that an instance belongs to a specific category.

→ Logistic regression finds a fit function f(x) such that,

P(y[i] in class) ~ f(x[i,]) = S(a + b[1] x[i,1] + ···· + b[n] x[i,n])

S[z] - sigmoid function. S(z) = 1/(1 + exp(z))

y[i] are the probabilities that x[i,] belongs to the class of interest.

→ Logistic regression is similar to linear regression that finds the log-odds of the probability that you are interested in.

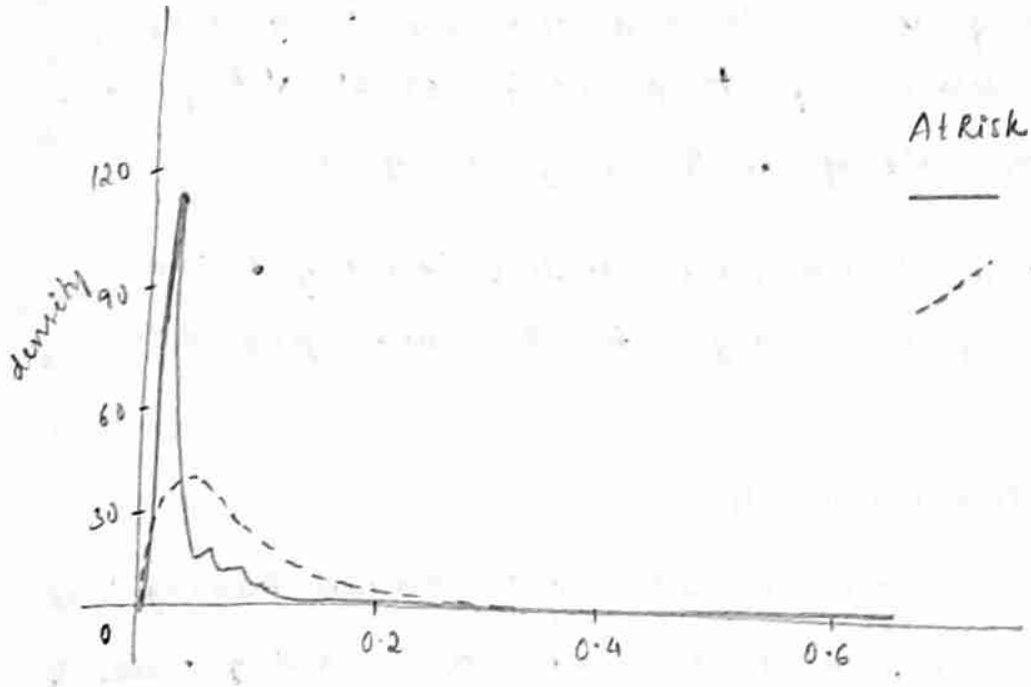
→ Logistic regression assumes that logit(y) is linear in values of x. Like linear regression, logistic regression will find the best coefficients to predict y, including finding advantageous combinations & cancellations when the inputs are correlated.

Characterizing Prediction quality.

If our goal is to use the model to classify new instances into one of two categories, then we want the model to give high scores to positive instances & low scores otherwise.

This can be checked by plotting the distribution of scores for both the positive & negative instances.

For eg:- Distribution of score broken up by +ve examples (TRUE) & negative examples (FALSE).



→ If the score distribution of the +ve & -ve instances are seperated well, we can choose appropriate threshold between the two peaks.

In the above case, the two distributions are not well seperated, which indicates the model can't build a classifier that simultaneously achieves good recall & good precision.

→ The ratio of classifier's precision to the average rate of +ves is called the enrichment rate.

→ The higher the threshold, the more precise the classifier will be

→ Once, the threshold is picked, the resultant classifier can be evaluated by looking at the confusion matrix.

→ The coefficients of logistic regression model encode the relationships between the input variables & the output in a way similar to ~~to~~ ka

→ Pseudo - R-squared is a useful goodness-of-fit heuristic.

245-18-733-00

5: What is Predictive modeling? Discuss about evaluation of Predictive models.

Predictive modeling is a statistical technique using machine learning and data mining to predict & forecast likely future outcomes with the aid of historical & existing data.

→ A predictive model is not fixed, it is validated & revised regularly to incorporate changes in the underlying data.

Evaluation of Predictive models.

Evaluation metrics have correlation with machine learning tasks. Metrics like precision, recall are used for evaluating models & the tasks of classification, regression, ranking, clustering, topic modeling etc, all have different metrics.

Evaluating a model is very important step throughout the development of the model.

Evaluation methods in supervised learning falls under two categories: 1. classification
2. regression.

In classification, evaluation means to focus on the number of predictions that were classified correctly.

In regression, evaluation means to identify the error between the actual & prediction output.

2451-18-733-001

Model Evaluation Techniques.

Model evaluation is an integral part of model development evaluation. It helps to choose find the best model that represents our data.

There are 2 methods of evaluating models,

1. Hold-Out 2. Cross-Validation.

1. Hold-Out

In this method, the mostly large data set is divided into 3 subsets:

(i) the training set.

(ii) The validation set

(iii) the test set.

2. Cross-Validation.

Used when only limited amount of data is available, to achieve an unbiased estimate of model performance we use k-fold validation.

The data set is divided in to k subsets of equal size

The model is built k times, each time leaving out one of subsets from training & use it as test set.

If k = sample size, then it is leave one out method.

Regression Model Evaluation methods.

1. Root Mean Square Error (RMSE)

Used to measure the errors rate of a regression model. It can be used to compare between two models whose errors are in same units.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}}$$

2. Relative Square Error (RSE)

can be used to compare between models whose errors are in different units.

$$RSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (\bar{a} - a_i)^2}$$

3. Mean Absolute Error (MAE)

It is the average of difference between the original values & Predictive values. They give measure of how far the predictions are from actual output but doesn't tell about direction of error.

$$MAE = \frac{1}{N} \sum_{j=1}^{N} |y_j - \hat{y}_j|$$

4. Relative Absolute Error (RAE)

can be compared between models whose errors are measured in different units.

$$RAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |\bar{a} - a_i|}$$

5. Coefficient of Determination (R^2)

R^2 summarises the explanatory power of regression model & is computed from the sum of squared terms

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Sum of squares total $SST = \Sigma (y - \bar{y})^2$

Sum of squares regression $SSR = \Sigma (y' - \bar{y}')^2$

Sum of squares error $SSE = \Sigma (y - y')^2$

R^2 describes the proportion of variance of dependant variable explained by the regression model.

If regression model is perfect then $SSE = 0$ $R^2 = 1$

If regression model is failure $SSE = SST$

no variance is explained & $R^2 = 0$

6. Standardized Residuals (Errors) Plot

The standardized residual plot is useful visualization tool in order to show the residual dispersion patterns on standardized scale.

There is no substantial differences between the pattern for a standardized residual plot & the pattern in the regular residual plot. The only difference is the standardized scale on the y-axis which allows us to easily detect potential outliers.

Classification Model Evaluation methods

1. Confusion matrix

it shows the number of correct & incorrect predictions made by the classification model compared to the actual outcomes in the data.

Confusion matrix :

		Target: Positive	Target: Negative
Model	Positive	a	b
	Negative	c	d

(1) Accuracy : proportion of total no. of on predictions that were correc

$(a+d)/a+b+c+d.$

(ii) Positive Predictive Value (Precision):

$$a/(a+b)$$

(iii) Negative Predictive value

$$d/(c+d)$$

(iv) Sensitivity or Recall

$$a/(a+c)$$

(v) Specificity:

$$d/(b+d)$$

2) Gain & Lift charts.

Gain or Lift

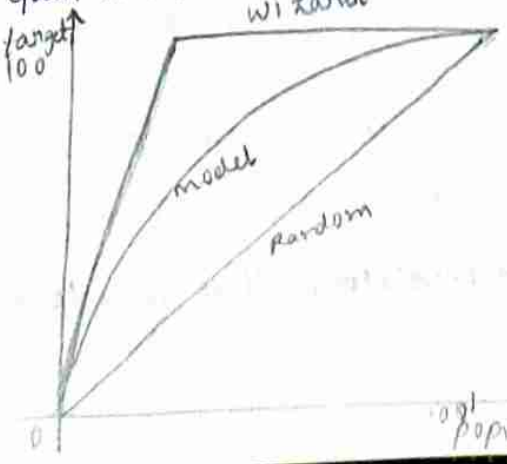
It is a measure of effectiveness of a classification model calculated as the ratio between the results obtained with & without the model.

These charts are visual aids for evaluating the performance of classification models.

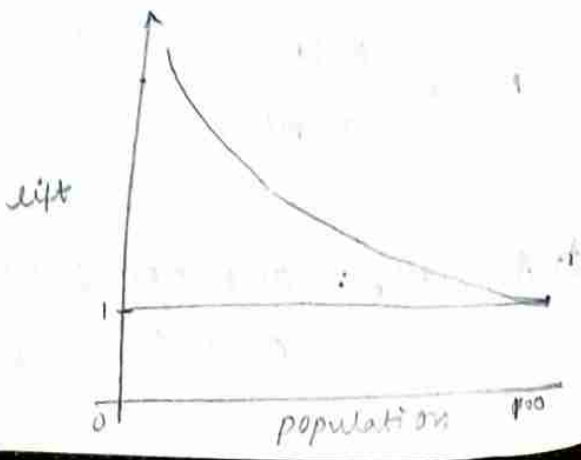
Gain & lift charts evaluates model performance in a portion of population.

The lift chart shows how much more likely we are to receive positive responses than if we contact a random sample of customers.

Gain chart



Lift chart



3. K-S chart

Kolmogorov - Smirnov chart measures the performance of classification models.

$\frac{(K-S)}{ft}$ is the measure of the degree of seperation between positive & negative directions.

4. ROC chart

It is similar to Gain or Lifts charts

ROC chart shows false positive rate (1- specificity) on X axis against true positive rate (sensitivity) on Y axis

Ideally, the curve climbs quickly toward the top left meaning the model has correct predictions.

5. Area under curve (AOC)

The area under ROC curve is often a measure of the quality of the classification model.

AUC for perfect classifier = 1

most of classifiers have AUC between 0.5 & 1.

6. Why logistic regression is used for classification. Explain model building strategies for logistic regression.

Logistic regression is used to in statistical software to understand the relationship between the dependent variable & one or more independent variables by estimating probabilities using a logistic regression equation.

This type of analysis can help you predict the likelihood of an event happening or a choice being made.

Logistic regression models helps us understand relationships and predict outcomes, you can act to improve decision-making.

There are varieties of model building strategies such as purposeful selection of variables, stepwise selection & best subsets.

The principal of model building is to select as less variables as possible ~~even though~~ even although the model reflects the true outcomes of data.

Steps involved in purposeful selection of variables.

1. Univariable analysis
2. Multivariable model comparisions.
3. Linearity assumption
4. Interactions among covariates
5. Assessing fit of the model

2. Stepwise selection of Covariates.

This method is used when the outcome being studied is relatively new & the important covariates may not be known & associations with the outcome may not be understood well. In these instances, most studies collect many possible covareates & screen them for significant associations. Employing a stepwise selection of covariates can provide a fast & effective means to screen a large number of variables, & to fit a number of logistic regression equations simultaneously.

→ Any stepwise procedure for selection or deletion of variables from a model is based on statistical algorithm that checks for the importance of variables & either includes or excludes them on the basis of a fixed decision rule.

→ importance of variable is defined in terms of measure of statistical significance of the coefficient for the variable.

3. Best Subsets Logistic Regression.

An alternative to best subset stepwise selection of a model is best subset selection.

Software implementing this method for linear regression identifies a specific number of 'best' models containing one, two, 3 variables & so on. up to single model containing all p variables.