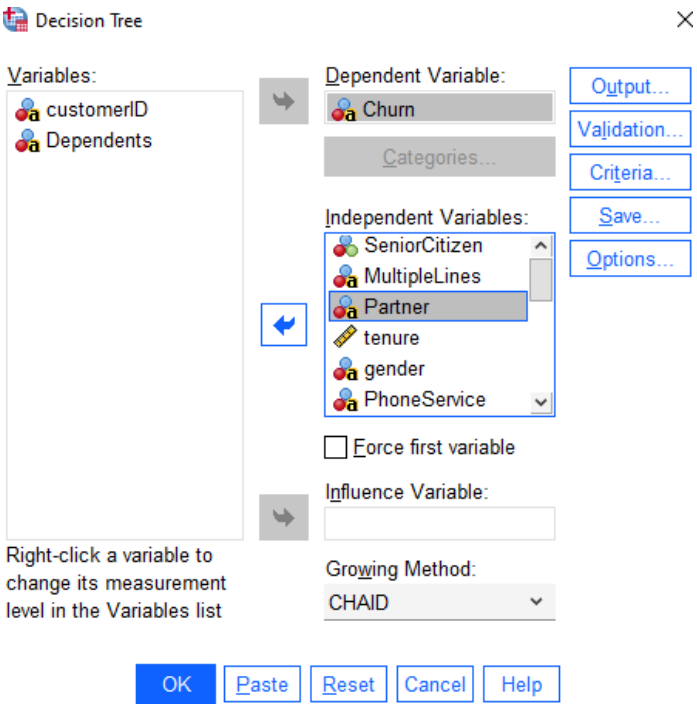
CSE 458: BUSINESS ANALYTICS – CA4

SHRIRAM KP
E0219007

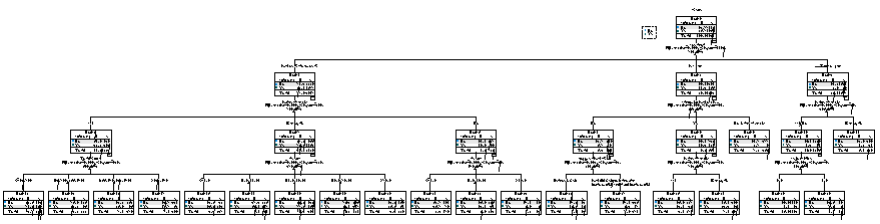
a) CLASSIFICATION

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The dependent and the independent variables are selected.



Model Summary

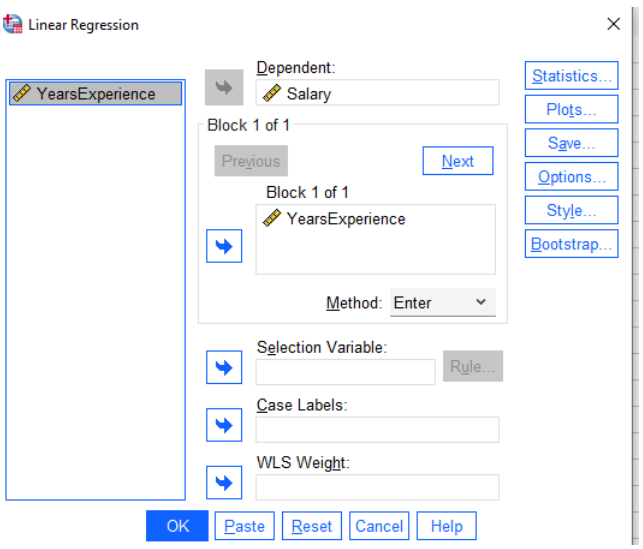
Specifications	Growing Method	CHAID
	Dependent Variable	Churn
	Independent Variables	SeniorCitizen, MultipleLines, Partner, tenure, gender, PhoneService, OnlineSecurity, TechSupport, DeviceProtection, InternetService, PaymentMethod, StreamingMovies, TotalCharges, PaperlessBilling, OnlineBackup, MonthlyCharges, Contract, StreamingTV
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Contract, InternetService, TotalCharges, tenure, StreamingMovies, PaymentMethod, SeniorCitizen
	Number of Nodes	30
	Number of Terminal Nodes	20
	Depth	3



The decision tree diagram is generated for each iteration and the depth of the tree is set to automatic.

b) REGRESSION

It is a simple dataset consist of a dependent and independent variable. The dependent variable is salary. The objective is to predict the salary using the years of experience. Once the dataset is imported, analyze > Linear regression is selected. Then, the dependent and independent variables is chosen in the tool



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.978[a]	.957	.955	5788.31505

a. Predictors: (Constant), YearsExperience

b. Dependent Variable: Salary

The larger the F-statistic, the greater the evidence that there is a difference between the group means and this case it is around 622 from the anova table. The residuals / error is lesser from the initial state as it can be seen in the Residuals Statistics and Coefficients table.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.086E+10	1	2.086E+10	622.507	<.001[b]
	Residual	938128551.7	28	33504591.13		
	Total	2.179E+10	29			

a. Dependent Variable: Salary

b. Predictors: (Constant), YearsExperience

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	25792.200	2273.053		11.347	<.001
	YearsExperience	9449.962	378.755	.978	24.950	<.001

a. Dependent Variable: Salary

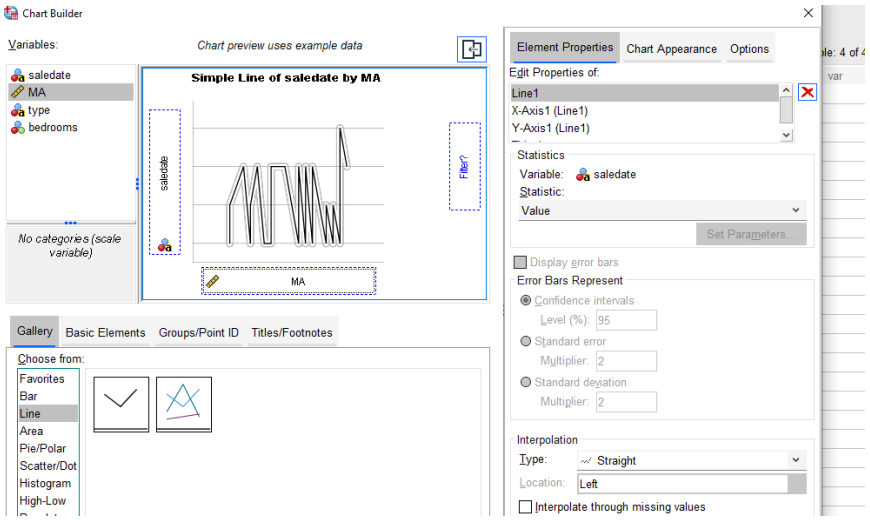
Residuals Statistics[a]

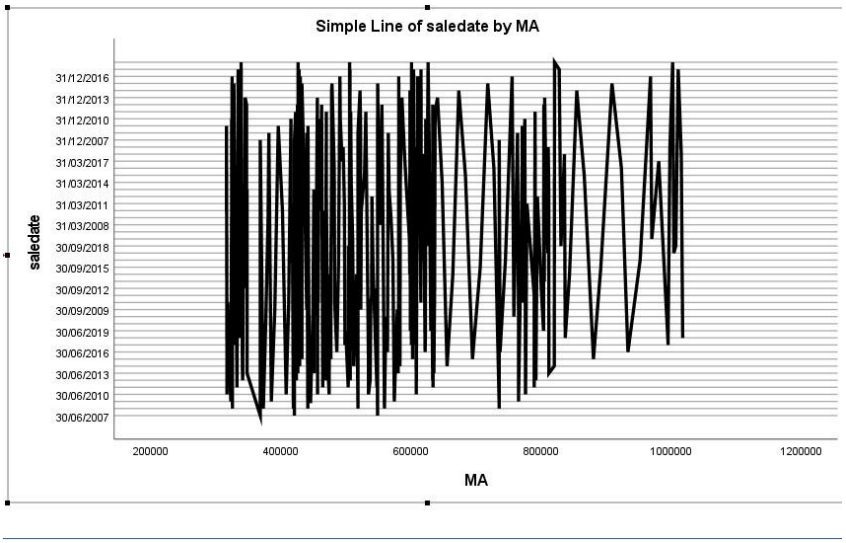
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	36187.1602	125016.8047	76003.0000	26817.93616	30
Residual	-7958.00781	11448.02539	.00000	5687.64102	30
Std. Predicted Value	-1.485	1.828	.000	1.000	30
Std. Residual	-1.375	1.978	.000	.983	30

The scatter plot points between the dependent variable and standardized predicted variable shows far the predicted and the actual values, it also shows the linearity.

c) FORECASTING

Time Series Forecasting predicts future values of a particular quantity based on previously observed values of that quantity. The chart builder option is selected and then x & y variable is chosen accordingly.





Plot shows the trend of moving average of median price throughout the years

Risk

Estimate	Std. Error
.204	.005

Growing Method: CHAID
Dependent Variable: Churn

Classification

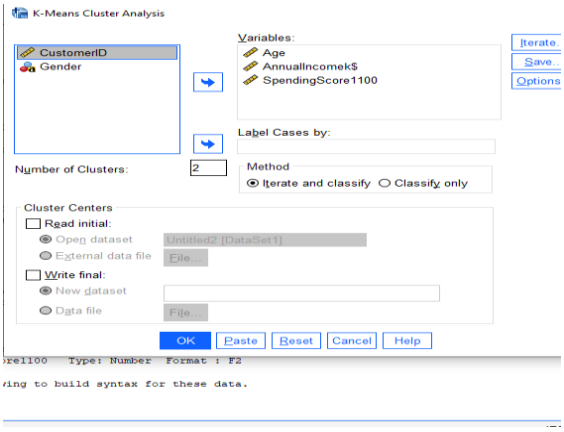
Observed	Predicted No	Predicted Yes	Percent Correct
No	4678	496	90.4%
Yes	943	926	49.5%
Overall Percentage	79.8%	20.2%	79.6%

Growing Method: CHAID
Dependent Variable: Churn

The model is able to do the classification with smaller error rate and the overall accuracy is approximately 80%.

a) CLUSTERING

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. To do the K – Means Clustering under the analyze tab , Classify > K Means is selected. Then the variables are selected to do the analysis.



Initial Cluster Centers

	Cluster	
	1	2
Age	64	30
AnnualIncomek$	19	137
SpendingScore1100	3	83

The centers and the initial cluster history is provided.

Iteration History[a]

Iteration	Change in Cluster Centers 1	2
1	56.164	50.032
2	.513	1.701
3	.370	1.201
4	.367	1.177
5	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 5. The minimum distance between initial centers is 146.561.

How the centers change throughout various is provided, the maximum is around 56 and the minimum 0.

Final Cluster Centers

	Cluster 1	2
Age	40	34
AnnualIncomek$	51	91
SpendingScore1100	44	71

ANOVA

	Cluster Mean Square	df	Error Mean Square	df	F	Sig.
Age	1397.633	1	189.060	198	7.393	.007
AnnualIncomek$	59003.319	1	395.323	198	149.253	<.001
SpendingScore1100	27324.632	1	532.219	198	51.341	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The final clusters show that the customers have been segmented into 2 different groups based on the centers and from the anova table it can be seen that the mean square is cluster mean square is more than the mean square of the individual columns.