CIDM/ECON 6308 Seminar in Data Analytics Exam 2

200 points; Due 11:59 CST April 23, 2023

Requirements: This exam is open book, open slides, and open notes. Because this is an individual exam, you are not allowed to collaborate nor discuss with anyone else during the exam period. Sharing anything related to the exam with any person or party violates the University's Academic Integrity Code, as well as the PEV COB Student Code of Ethics listed in our syllabus, and will be reported to the Dean Office of PEV COB. Any question about the exam should be directed to the instructor.

Please follow the instruction and requirements to answer each question. After completing the exam, please submit your screenshots and response to open-ended questions via Exam 2 Screenshot Submission while your answers to other questions via Exam 2 Objective Questions Submission on WT Class. It is your responsibility to make your screenshots and answers meet the required format; otherwise, you would lose points because of wrong format. Please use Exam 2 Screenshot Submission Template to compile your screenshots and responses and then submit it on WTCLASS. If you do not use it as a template, your submission WILL NOT be graded and a zero point will be assigned.

Exam 2 is designed to help you solve real-world business problems using data analytics techniques that you have learned in this semester, which will help you prepare for your group project. It includes five tasks below:

- Task 1: Understanding Analytics Concepts & Principles (20 points)
- Task 2: Demonstrating the Importance of Retaining Customers (30 points)
- Task 3: Data Preparation & Exploration (36 points)
- Task 4: Building & Understanding Decision Tree Model in RapidMiner (30 points)
- Task 5: Prediction with Decision Tree and Logistic Regression Models and Model Comparison (54 points)
- Task 6: Time Series Decomposition (30 points)

Datasets:

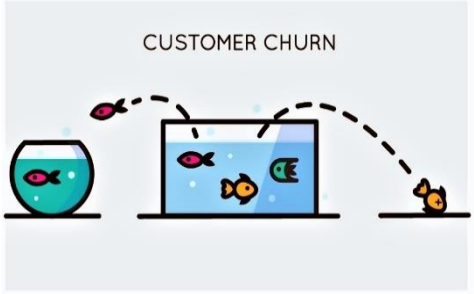
- Training.csv, the training set with 5,000 records, used for Task 2, 3, 4, and 5.
- Prediction.csv, the prediction set with 100 records, used for Task 5.
- Revenue.csv, the time series with quarterly revenue from 2004 to 2020, used for Task 6.

Software:

- Excel
- Tableau
- RapidMiner

Our textbook *Data Science for Business* mentions the example of predicting customer churn (a typical application of data analytics) in multiple chapters such as Chapters 1, 3, 5, 8, and 14.

Mark just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States. MegaTelCo provides both wireless and internet services and it has hundreds of millions of customers. They are having a major problem with customer retention in their telecommunication business. Many customers leave, and it is getting increasingly difficult to acquire new customers. Since the telecommunication market is now saturated, the huge growth in the telecommunication market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own. Customers switching from one company to another is called customer churn, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs. According to a report, annual churn rates for telecommunications companies average between 10% and 67% (Database Marketing Institute, 2008). Customer churn not only increases operation and advertising cost, but also reduces revenue and damages brand image. As Computer Weekly cites, mobile operators spend approximately 15 percent of their revenues on network infrastructure and IT — but a whopping 15 to 20 percent of revenues on the acquisition and retention of customers (Computers Weekly, 2018).



It's long been known retention of existing customers is less expensive than acquisition of new ones. In fact, a Canadian study found it costs nearly 50 times less to retain than acquire (Telecoms, 2018). Therefore, a good deal of marketing budget is allocated to prevent customer churn. The Marketing department is going to designate a special retention offer. Mark's task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to predict whether or not a particular customer is going to turn over before s/he actually leave so that MegaTelCo can offer the special retention deal to prevent customer churn. This is even more important to retain high-value customers. In order to predict customer churn, Mark and his data science team need to apply data analytics techniques (esp., data mining). In order to solve this problem, Mark plans to take the data on prior churn and extract patterns, for example, patterns of behavior, that are useful—that can help him to predict those customers who are more likely to leave in the future, or that can help us to design better services.

Considering that it is very time-consuming to download and process millions of records, Mark decides to start with building data mining models via a portion of the data via a random sampling technique (See Chapter 8 in Dr. B's book). Using the database querying technique (to be covered in Class 06), Mark obtains a random sample of 5000 records (i.e., customers) from the company's data repository.

Next, Mark prepares the data for initial analysis: First, as the dataset has hundreds of attributes, Mark applies feature selection techniques to include a small number of important attributes in his initial models, rather than all the attributes. Next, Mark cleans the data to solve the quality issues of the data such as missing or extreme values. Finally, Mark obtains a cleaned dataset with 13 predictor attributes (or variables) from three categories (see the table as below) and one target attribute (i.e., the attribute of our interest, also called dependent attribute in statistics).

Attributes and Their Description

Category	Attribute Name	Description	Values
Demographic Information	CustomerID	A Unique ID to identify each customer	Format: NNNN-LLLLL
	Gender	The gender of each customer	Male/Female
	SeniorCitizen	Whether or not a customer was a senior citizen	1=yes, 0=no
	Partner	Whether or not a customer had a partner	Yes/No
	Dependents	Whether or not a customer had dependent(s)	Yes/No
	Tenure	The length of time (in months) a customer had stayed with the company	Whole number
Service Information	PhoneService	Whether or not a customer had phone service	Yes/No
	InternetService	What type of internet service a customer had (code No if a customer had no internet service)	DSL/Fiber/No
Account Information	Contract	The contract type that a customer had with the company	Month-to-Month/ One Year/ Two Year
	PaperlessBilling	Whether or not a customer used paperless billing	Yes/No
	PaymentMethod	The payment method that a customer used	Electronic Check Mailed Check credit Card (auto) Bank Transfer (auto)
	MonthlyCharges	The total monthly payment a customer made	Real number
	TotalCharges	The total life-to-date value/revenue a customer contributed	Real number
Target	Churn	Whether or not a customer churned last month	Yes/No

In Exam 1 Part 2, you helped Mark gain a good theoretical understanding of this case and explored the data and the relationship between other attributes and the target attribute, churn. In Exam 2, you are going to continue to help him work on this case using multiple data analytics techniques to accomplish the following four goals:

1) Demonstrating the importance of retaining customers
2) Preparing and exploring the data before modeling.
3) Building a decision model to predict whether or not a particular customer is going to turn over and presenting this model to the executive team.
4) Making predictions with decision tree and logistic regression models and comparing their results.

In order to achieve the goals above, you are going to apply data analytics techniques that you have learned from the course Data Analytics Seminar at WTAMU.

1 Understanding Analytics Concepts & Principles (20 points)

In the past few weeks, we have introduced a list of important analytics methods, concepts, and principles. Please indicate whether each of the ten statements on WTCLASS is true or false by typing T or F in the front (20 points: 2 points for each question).

2 Demonstrating the Importance of Retaining Customers (30 points)

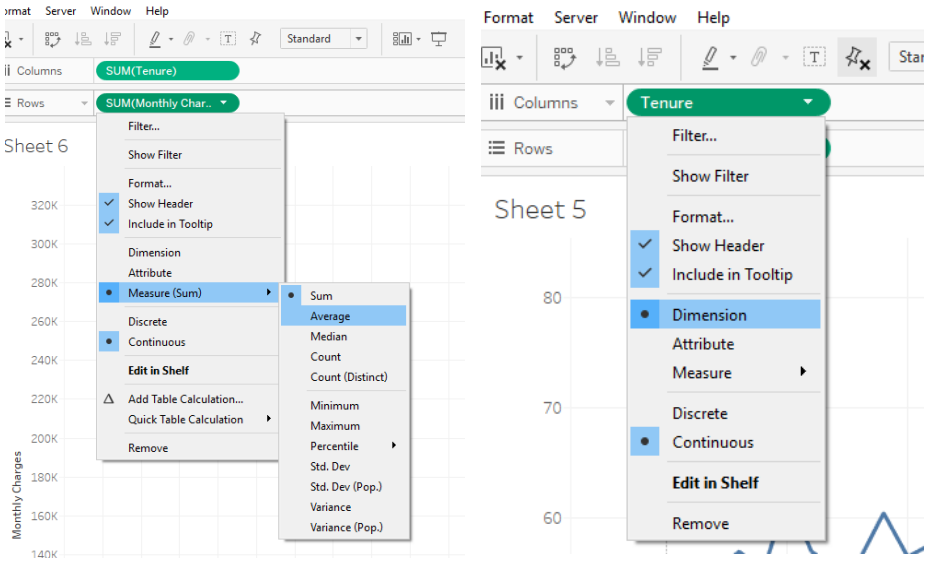
In order to demonstrate the Importance of Retaining Customers, you want to use the attribute Tenure in your dataset. It is well-known that a customer with a longer tenure is more loyal and contributes more total revenue to the company. In this case, you want to examine whether customers with a longer tenure pay more each month. Here, you are not analyzing each individual customer, but customers at each tenure level. In Tableau, you are going to generate a line chart to show the relationship between tenure and the average monthly charges of all the customers at each tenure level.

2.1 Import the data into Tableau.

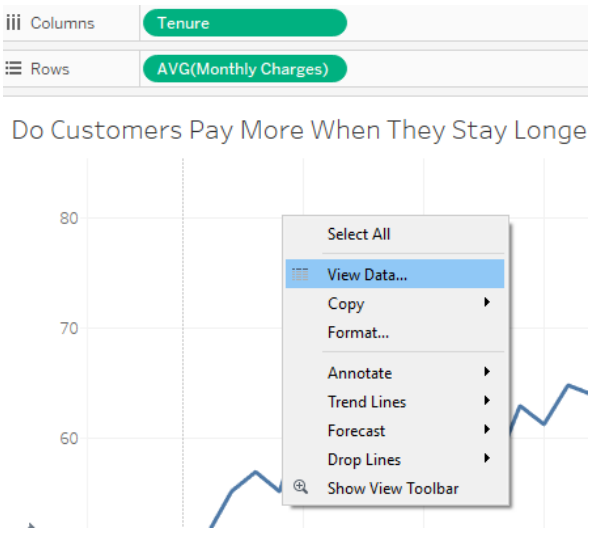
2.2 Drag Tenure and MonthlyCharges to the Columns and Rows, respectively.

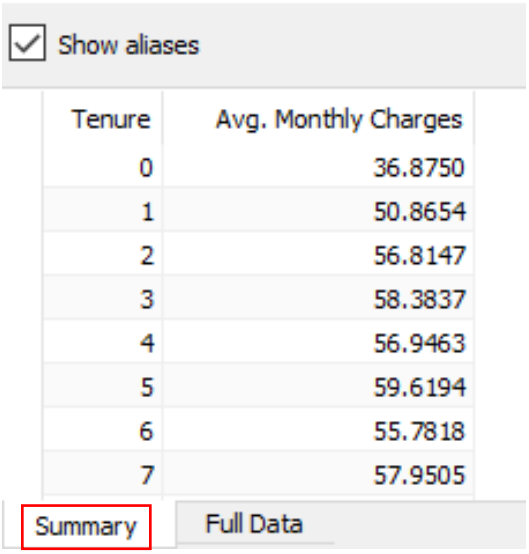
2.3 Change the measure of Monthly Charges to Average

2.4 Change Tenure to Dimension and then you will see a line chart



2.5 This line chart shows the pattern between tenure and the average monthly charges of all the customers in each tenure period. Overall, the line chart has an increasing trend with some fluctuation. Click "View Data" and you will see the dataset displayed by the line chart (Note: your dataset may be different from the one provided below as we generate a different random sample each time).





2.1 As you see, the dataset is displayed in Summary does not have 5,000 records because customers at each tenure level (e.g., 6 months) are cumulated/summarized into one data point to represent their average monthly charges. Even though the data here has a smaller granularity (less granular), it can help us answer the question whether or not customers would pay more as they stay longer in a company. View the data (Summary) and the line chart, and then answer the following questions (6 points in total, 2 points for each blank; type whole numbers only):

2.1.1 How many records does the dataset (Summary) have? _____

2.1.2 The Avg. Monthly Charges is greatest when Tenure = _____

2.1.3 The Avg. Monthly Charges is smallest when Tenure = _____

2.2 Click Analytics pane and drag the default Trend Line to your line chart (the default trendline in Tableau a linear regression model).

2.3 When moving your cursor to your trendline, you will see your trendline model. Based on the trendline model and what you learned in Class 10, answer the following questions (12 points in total and 2 points for each blank). Round your answers to the second decimal place such as 0.12.

2.3.1 You see a regression model with the trendline. Please complete the model as below:
Average Monthly Charges = ____× Tenure + _____

2.3.2 This linear regression model suggests that when a customer's tenure increases by five months, their monthly charge will ____(type increase or decrease) by ____ dollars.

2.3.3 You find R^2 (R Squared) = ____

2.3.4 According to Class 10 Lecture, R^2 is important measure of the goodness of fit of a linear regression model. In simple linear regression (there is only one independent variable), the square of the correlation between the dependent variable and the explanatory variable. Please use R^2 to compute the correlation coefficient between tenure and average monthly charges. Their correlation coefficient = _____.

2.3.5 Finally, the p-value indicates that this regression model is quite significant, based on what you have learned from Class 10 Lab.

2.4 Take a screenshot of your line chart with the trend line with date and time (Screenshot 1). Your screenshot must clearly include a line chart, a trendline, x axis with a meaningful label, and y-axis with a meaningful label, and display date and time. Then, answer the questions: What this chart is about? What meaningful pattern or rule can you observe from the chart? Based on what you observe, please present why it is important to retain customers (12 points: 5 points for your screenshot and 7 points for your responses; the detailed requirements for your responses are specified in the Screenshot Submission Template).

3 Data Preparation & Exploration (36 points)

3.1 There are three numerical attributes in our dataset, Tenure, Monthly Charges, and Total Charges. We are going to see if there is any redundant or highly-correlated attribute there. Please compute the correlation coefficients between any two of the three attributes (8 points in total and 2 points for each blank). Round your answers to the second decimal place such as 0.34.

- The correlation coefficient between Tenure and Monthly Charges is _____ (Note: This correlation is different from the one in Step 2.3.3 because they have different units of analysis: UOA in Step 2.3.3 is customer group while here the UOA is individual customers).
- The correlation coefficient between Tenure and Total Charges is _____
- The correlation coefficient between Monthly Charges and Total Charges is _____

- Two attributes are highly correlated when the absolute value of their correlation coefficient is greater 0.85. In this case, how many pairs of attributes are highly correlated? Type a whole number here. ___

3.2 Mark wants to explore how many clusters those customers naturally form using the k-means clustering. He tries to figure out how k-means clustering works before running it on RM. The two variables, Tenure and Monthly Charges, are used for clustering customers into four clusters. After a particular iteration, the centroid of each cluster is computed below (Tenure, Monthly Charges):

- Cluster 1 (11,32): Tenure =11 months, Monthly Charges =$32
- Cluster 2 (52, 33): Tenure =52 months, Monthly Charges =$33
- Cluster 3 (20, 81): Tenure =20 months, Monthly Charges =$81
- Cluster 4 (64, 93): Tenure =64 months, Monthly Charges =$93

Next, he needs to compute the distance from each data point (i.e., a customer) to each centroid above to determine to which cluster each data point belongs. In order to illustrate k-means algorithms to others, he uses a customer (i.e., a data point) with a tenure of 38 months and $95 in monthly charges (38, 95) as an example (20 points).

3.2.1 He first computes the Manhattan distance from this data point to each centroid. Type a whole number in each blank.

- The Manhattan distance from this specific data point to the centroid of Cluster 1 is 90.
- The Manhattan distance from this specific data point to the centroid of Cluster 2 is _____.
- The Manhattan distance from this specific data point to the centroid of Cluster 3 is _____.
- The Manhattan distance from this specific data point to the centroid of Cluster 4 is _____.
- Based on the four distances above, this specific customer will be assigned to Cluster _____. Type 1, 2, 3, or 4 here.

3.2.2 He then computes the Euclidian distance from this data point to each centroid. Round each distance to a whole number.

- The Euclidian distance from this specific data point to the centroid of Cluster 1 is _____.
- The Euclidian distance from this specific data point to the centroid of Cluster 2 is _____.
- The Euclidian distance from this specific data point to the centroid of Cluster 3 is _____.
- The Euclidian distance from this specific data point to the centroid of Cluster 4 is _____.
- Based on the four distances above, this specific customer will be assigned to Cluster _____. Type 1, 2, 3, or 4 here.

3.2.3 Do the two distance measures (Manhattan distance and Euclidian distance) assign the same cluster to this data point? Type Yes or No here.

3.3 Mark also explores that there are many networks among those customers based on their phone call records. After doing some research, he realizes that network measures such as density and centrality can also influence customer churn. In order to understand those measures, he explores four undirected networks below (8 points).

- Network 1: 26 customers and 180 connections based on their phone call records
- Network 2: 23 customers and 135 connections based on their phone call records
- Network 3: 20 customers and 120 connections based on their phone call records
- Network 4: 17 customers and 90 connections based on their phone call records

Please help him answer the following questions:

- Which network has the largest density (Type 1, 2, 3, or 4)?
- What is its density score? Round your answer to the second decimal place such as 0.12
- Which network has the smallest density (Type 1, 2, 3, or 4)?
- What is its density score? [a4] Round your answer to the second decimal place such as 0.12.

4 Building & Understanding Decision Tree Model in RapidMiner (30 points)

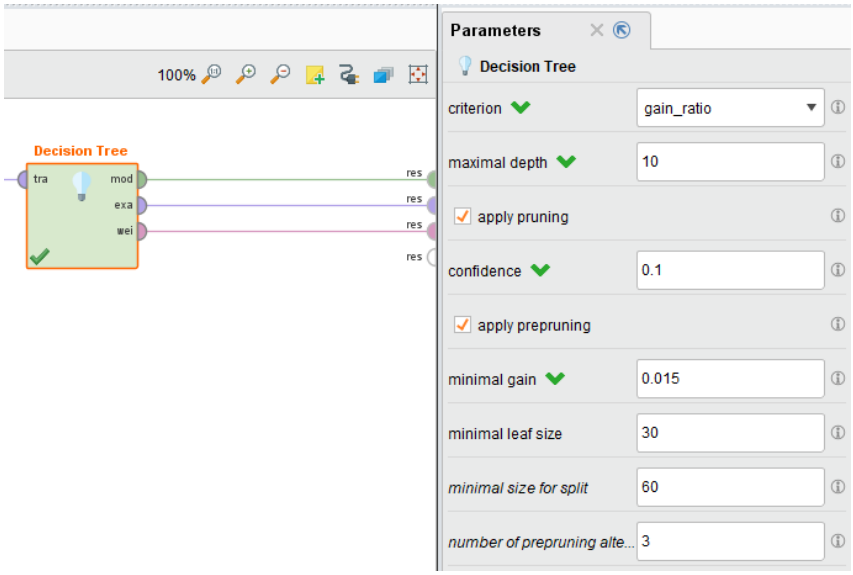
4.1 Import the data to RapidMiner (You may follow Class 10 Lab).

4.2 Unselect highly-correlated (see Step 3.1) or irrelevant (e.g., Customer ID) attributes using an operator in RapidMiner (you may refer to Class 9 Lab for this operator). You must remove those attributes in RapidMiner using this operator, instead of manually removing those attributes in Excel.

4.3 Set your target attribute.

4.4 Develop a decision tree model using the specified parameters below.

4.5 Here, you are requested to provide three outputs: model, example set, and weights. We learned the first two outputs in our previous lab. The third output (weights) represents the feature importance for each given attribute. A weight is given by the sum of improvements the selection of a given attribute provided at a node. The amount of improvement is dependent on the chosen criterion. The higher the weight of a given attribute, the more important or informative it is in the decision tree model. This can be used to identify the strong differentiators that we discussed in Exam 1 Part 2.



4.6 Save your RM process and then run it.

4.7 After running the process, please take a screenshot of your decision tree model (the one with a tree graph) with date and time (Screenshots 2). Then, present the decision tree to someone such as Mark's boss who have no background about data mining (14 points: 5 points for your screenshot and 9 points for your presentation; the detailed requirements for your presentation are specified in the Screenshot Submission Template).

4.8 Please answer the following questions (10 points in total and 2 point for each). Type a whole number in each blank unless otherwise stated.

4.8.1 Excluding the root node, how many split nodes in this tree?_____

4.8.2 Among all the leaf nodes, _____ are labeled as Yes and _____ are labeled as No.

4.8.3 Among all the leaf nodes, how many of them is 100% pure (i.e., single color in the leaf node)? ___

4.9 Sort Attribute Weight Output by the weight in descending order and then answer the following questions (6 points in total and 2 points for each):

4.9.1 Which attribute has the highest weight? Type the attribute name here. _____

4.9.2 The weight of this attribute is _____ (Round to the second decimal place).

4.9.3 Is this attribute also the root node in your decision tree graph? Type Yes or No here. _____

5 Prediction with Decision Tree and Logistic Regression Models and Model Comparison (54 points)

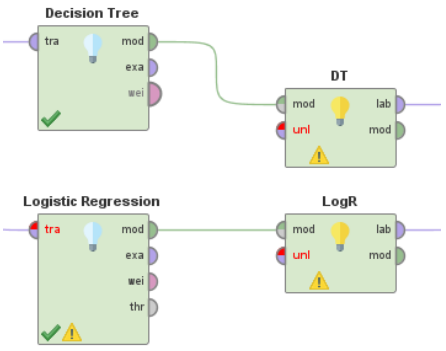
Now, you are asked to predict whether or not 100 new customers (stored in Prediciton.csv) will churn using both decision tree and logistic regression models in **a** RM process. The Logistic Regression operator in the new version of RM is getting more powerful as it can automatically handle binomial string attribute *Churn* and recode nominal or categorical attributes to dummy attributes. Accordingly, you do not need to transform Yes and No to 1 and 0, respectively. If you do not know how to allow two operators to use the same dataset (training or prediction), you can refer to an example in the Appendix at page 11.

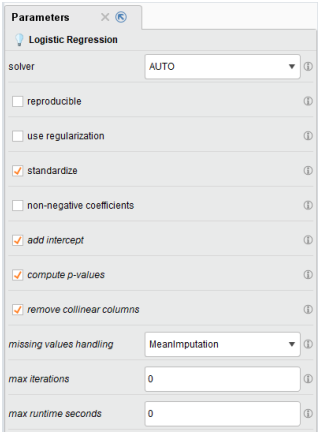
5.1 Use the decision tree model you built in Step 4 for prediction.

5.2 Develop a logistic regression model for prediction with the specified parameters of the Logistic Regression operator below (the specified parameters are default, but just in case that your default is different from mine, please double check).

5.3 In order to distinguish the prediction results of the two models, please change the name of Apply Model for the decision tree model and logistic regression model as DT and LogR respectively.

5.4 Run your process to generate two predictions results: one from the decision tree model and the other one from the logistic regression model.





5.5 Take a screenshot of your RM process with date and time (Screenshot 3). Requirements: your screenshot must include all the operators (both training and prediction steps, both decision tree and logistic regression models) in a correct sequence in ONE process. Then, please present the data mining process (CRISP-DM) using your RM process to someone who has no data mining background. (24 points: 6 points for your screenshot and 18 points for your presentation; the detailed requirements for your presentation are specified in the Screenshot Submission Template)

5.6 Take a look at the statistics view of the prediction results from your decision tree model and logistic regression model. An example from another project is provided below.

Note: The example below is used to help you understand the statistics of the three attributes: Prediction and two Confidence values. For the two confidence attributes, you can see how they are distributed and summary statistics such as min, max, or average, standard deviation. Ideally, we hope that the confidence, i.e., the estimated probability of being in any class, is equal or close to 0 or 1 (i.e., the two ends in the histogram). As shown in the two histograms above, almost all the confidence values are located at the both extremes, indicating that the overall prediction results are quite confident. In order to better evaluate and compare prediction performance of various models, you can either find online materials,

read our textbook (Chapters 7 and 8), and/or take CIDM 6355 Data Mining Methods (offered each Fall).

5.7 Based on the statistics views above, please answer the following questions (24 points in total and 4 points for each blank):

5.7.1 For decision tree model, _____ customers are predicted to churn (i.e., Churn = Yes). Type a whole number here.

5.7.2 For the decision tree model, the maximum confidence of not churning (i.e., confidence(No)) is _____ (round to the third decimal place).

5.7.3 For the decision tree model, the maximum confidence of churning (i.e., confidence(Yes)) is _____ (round to the third decimal place).

5.7.4 For logistic regression model, _____ customers are predicted to churn (i.e., Churn = Yes). Type a whole number here.

5.7.5 For the logistic regression model, the maximum confidence of not churning (i.e., confidence(No)) is _____ (round to the third decimal place).

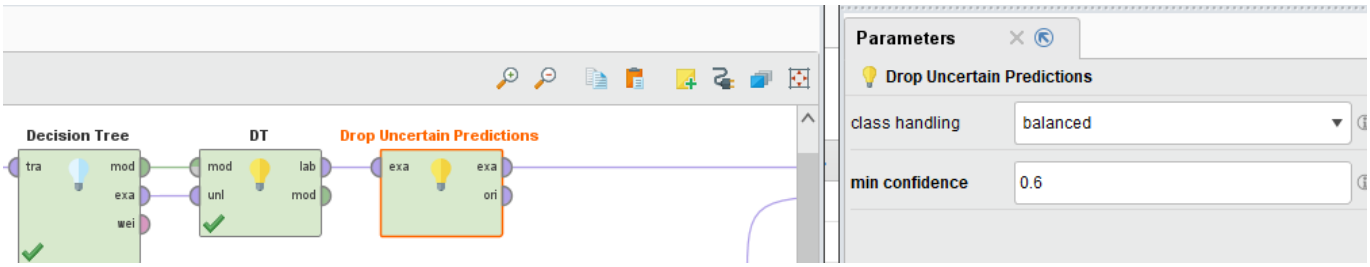
5.7.6 For the logistic regression model, the maximum confidence of churning (i.e., confidence(Yes)) is _____ (round to the third decimal place).

5.8 As mentioned in the note above, we try to avoid any uncertain predictions. In our case, the target attribute is a binary, Yes or No. Therefore, the confidence (Yes) = 1 – confidence (No). We decide to drop the prediction for those records with a confidence between 0.4 and 0.6 (excluding 0.4 and 0.6). Please answer the following questions (6 points in total and 3 points for each; Type whole numbers only):

5.8.1 How many predictions from the decision tree model will be dropped because of high uncertainty?

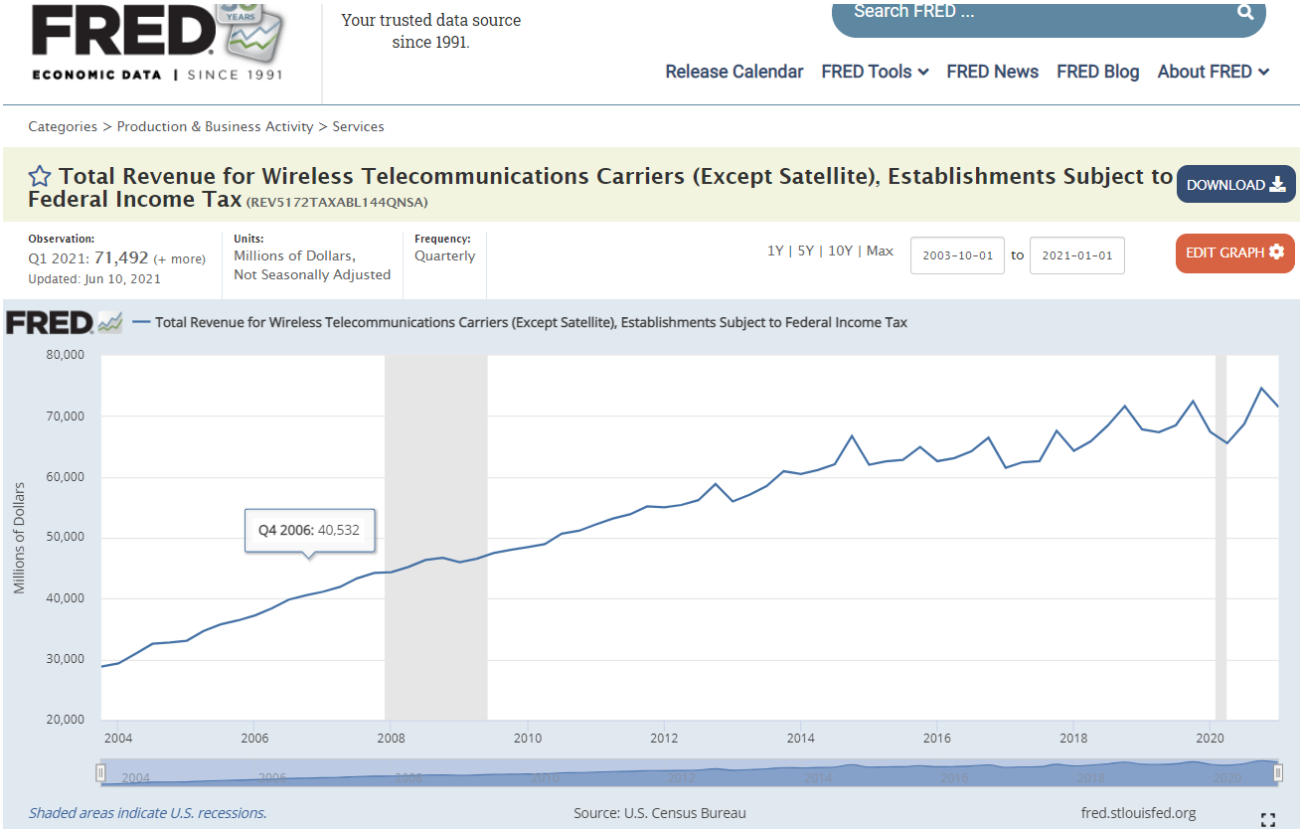
5.8.2 How many predictions from the logistic regression model will be dropped because of high uncertainty?

Note: You can either manually count the number by sorting confidence (Yes) or confidence (No) or add one more operator “Drop Uncertain Predictions” with the specified parameters below.



6 Time Series Decomposition (30 points)

Mark is also asked to understand the pattern of the total revenue in the wireless telecommunication industry, which will be used to predict the future revenue of the industry and that of MegaTelCo. The dataset Revenue.csv is obtained from FRED. This dataset describes quarterly total revenue for wireless telecommunications carriers (except satellite) in United States (million dollars) from 2004 to 2020. A time plot is provided below. You will help Mark decompose the time series.



Observing the time plot above, you find that the quarterly revenue has an upward trend, indicating that it increases over time. It also has seasonality, which is getting stronger since 2012.

6.1 When performing classical additive decomposition, which of the following moving averages is the estimate of the trend-cycle component? (3 points)

A. 2-MA
B. 4-MA
C. 8-MA
D. 2×4-MA
E. 3-MA
F. 4×4-MA

6.2 You can choose Excel or RM to decompose the time series using the classical additive method and then answer the following questions (27 points in total and 3 points for each blank):

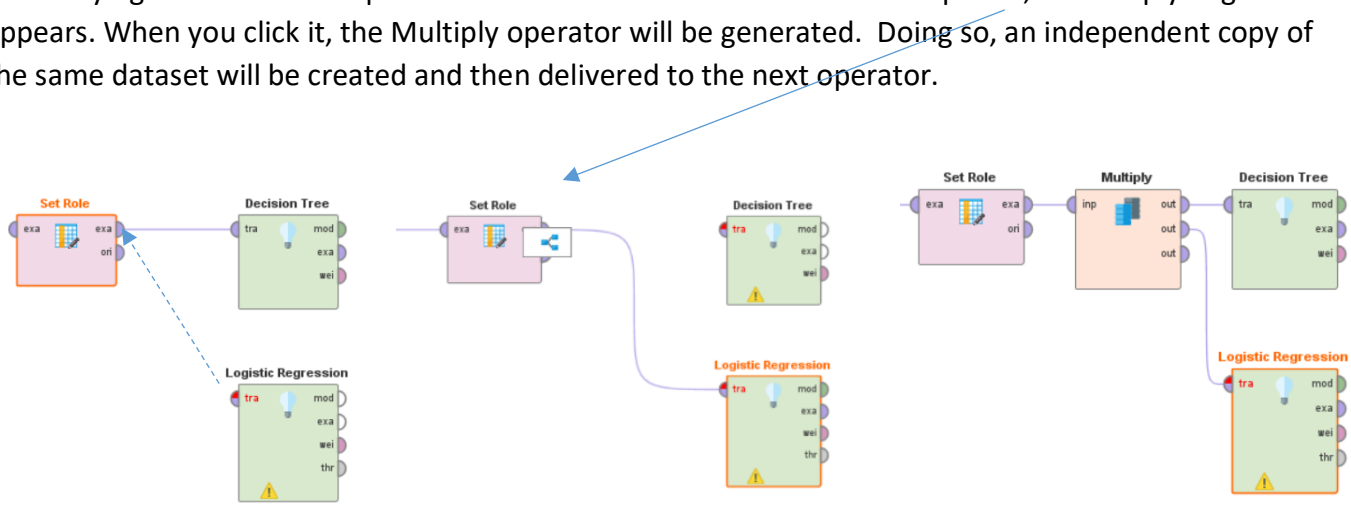
6.2.1 In total, the estimate of the trend-cycle component is available for ___ observations; the estimate of the seasonal component is available for ___ observations; the estimate of the remainder component is available for ___ observations (type a whole number for all the blanks).

6.2.2 The estimate of the first available trend-cycle component is _____; the estimate of the first available seasonal component _____; the estimate of the first available remainder component is _____; the estimate of the first available seasonally adjusted value is _____. (Round to whole numbers).

6.2.3 Based on the time plot and the decomposition result, generally, Quarter ____ has the highest revenue, while Quarter ____ has the lowest revenue. Type 1, 2, 3, or 4 here.

Appendix

When trying to connect the operator for the second model with the ExampleSet, a "multiply" sign appears. When you click it, the Multiply operator will be generated. Doing so, an independent copy of the same dataset will be created and then delivered to the next operator.



The Multiply operator takes the RapidMiner Object from the input port and delivers copies of it to the output ports. In this case, the example set is copied so that both Decision Tree and Logistic Regression can use the example set.