**MDS 560 Week 2 Hands-On Accelerator**

Your deliverables and hands-on activities for this week are:

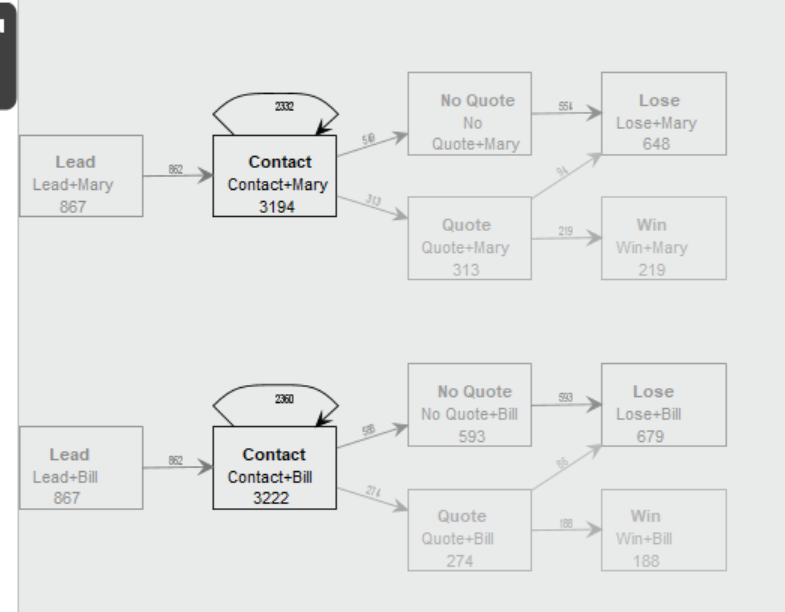
1. Compare and contrast the data mining CRISP-DM and the six sigma DMAIC methodologies. How are the two similar; how do they differ?

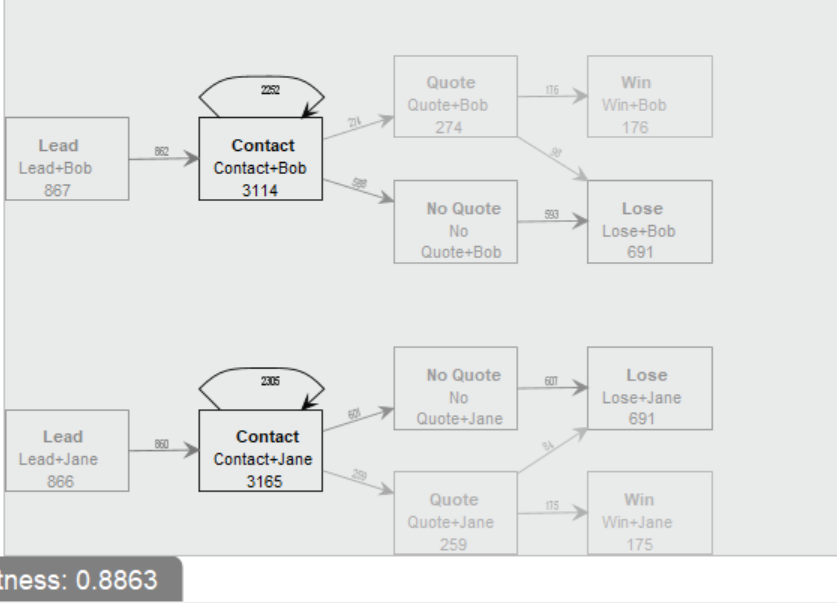
Results: The CRISP-DM methodology has 6 steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. DMAIC has 5 steps: Define, Measure, Analyze, Improve, and Control. Each methodology is iterative in nature, that is, the steps are repeated in effort to improve the process defined in the project objective. The two methodologies are similar in some ways such as, they both begin by trying to understand more about the process or business objective they are trying to improve. Business Understanding and Define steps from each respective methodology is the step in which the project team would dive deeper into process and define the overall objectives. Measure and Data Understanding/Data Preparation are all about gathering data and learning about the process from a data only view. What does the data tell us? Evaluation and Analyzing are similar because they both aim to interpret the results of the data in terms of the business objectives. While Analyzing might be reviewing the deliverables of the prior steps, Evaluation is interpreting figures provided by candidate modeling efforts, selecting the most appropriate model, and attempting to turn the results into actionable insights. Deployment is closely related to the Control phase where the actionable results of the project are implemented and closely monitored for a performance gain. At the same time, new questions and observations are being asked in order to look for opportunities to improve the process. The two methodologies contrast in the way that DMAIC is mostly applied to quality control projects. There are no modeling tools usually used, but a series of charts, maps, benchmarking, brainstorming, etc. activities. CRISP is more uniquely defined for a comprehensive modeling process that helps move from business objections to data gathering to modeling to evaluation/deployment in effort to meet the business requirements with actionable insights produced by a model. As discussed in lecture, mapping a process using event log data and applying some feature engineering concepts on process related data can help bring these two “worlds” together to help improve processes in a data mining approach.

2. Import Dataset 1b dataset into Rapidminer: Perform process discovery and analysis using Rapidprom operators (for Heuristics Miner) and assess process event conformance. Optionally perform a similar analysis using Pertri Nets, BPMN, Fuzzy Miner etc. Show your results.

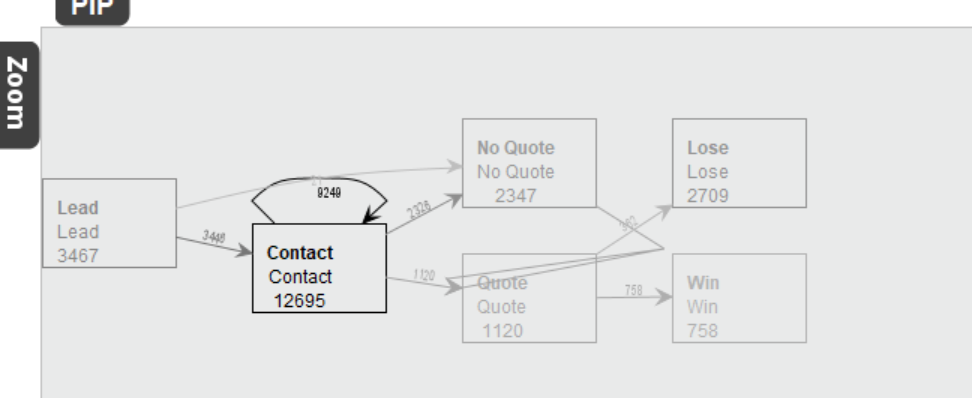
Results: I had to shorten the dataset in order to get it to run on my machine due to memory constraints. Also, I included screenshots in order to show my work. Please let me know if this is what you are asking for as a way to show work. Thanks.

Heuristics Miner: Activity and Resource

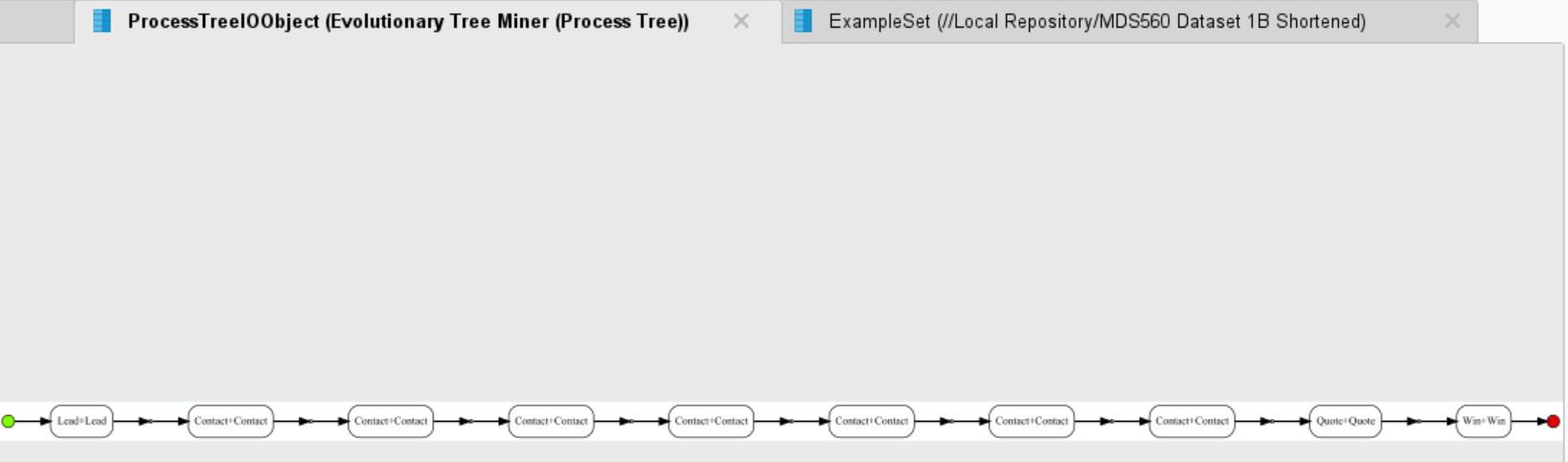




Heuristics Miner: Activity



Evolutionary Tree Miner



This says Lead/Lead + Contact/Contact(x7) + Quote/Quote+Yes/Yes.

I did both the Heuristic Miner and Evolutionary Tree Miner process discovery approaches. The Heuristic Miner made it pretty clear that leads that do not receive a quote, do not result in a win. Leads that do receive a quote are much more likely to be a win. This isn’t overly surprising information. How can a potential client purchase goods or services if they aren’t first quoted? My questions from there would include why aren’t these people being quoting? It doesn’t appear to be a sales representative in particular not quoting. They all appear to have similar numbers. In the Evolutionary Tree Miner, it provides an optimized process. It mentions contacting the potential customer 7 times before delivering a quote. This reminded me that maybe a representative gets shut down by the client before being able to establish a relationship with the client. Perhaps the sales representative isn’t doing an appropriate sales pitch that turns the client off from wanting to be reached further. Perhaps what turns these potential wins to losses are just not following up appropriately. Maybe their CRM should be alerting the sales rep to follow up sooner than what they are currently. I guess what I might first explore is if the number of times a customer is contacted is important rather then maybe the demographics of the customer.

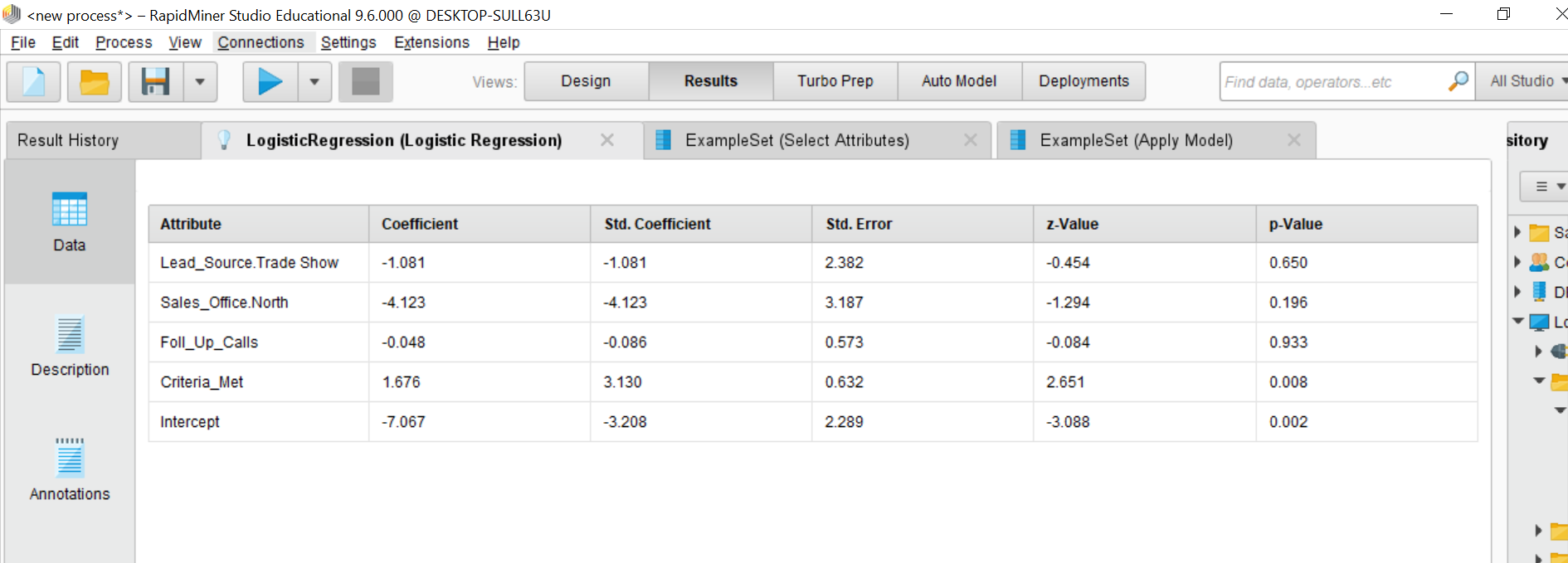
3. Watch Rapidminer video with customer churn example: <https://www.youtube.com/watch?v=qlW5Oo9TAsc>

Results: I found this video to be a very helpful introduction on what RapidMiner can do. It discusses the basic end to end data science steps of how you retrieve and clean data, apply models, use different validation techniques, optimize parameters, and make predictions to a test set. These are all steps discussed quite often in the modeling process. What hasn’t been discussed yet that I found really excited being in IT, is RapidMiner’s scheduling and web service features that allow for real deployment solutions and usability of the results.

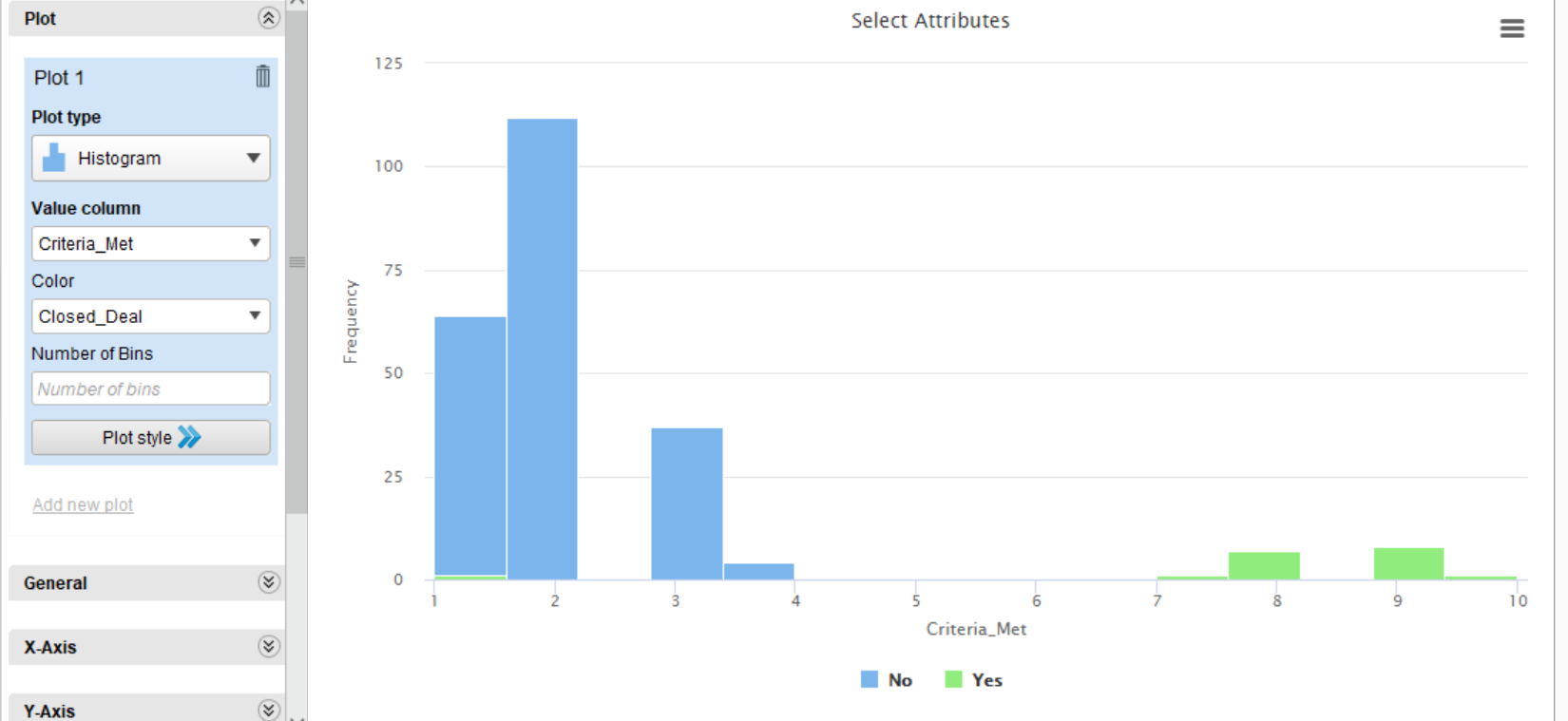
4. Import Dataset 2 dataset into Rapidminer. Perform Logistic Regression for classification modeling using “Closed\_Deal” as the label, and then run linear regression analysis using “Closed\_Flag” as the Y variable to explain, predict and verify sales performance using the given independent variables. Show your results.

Results:

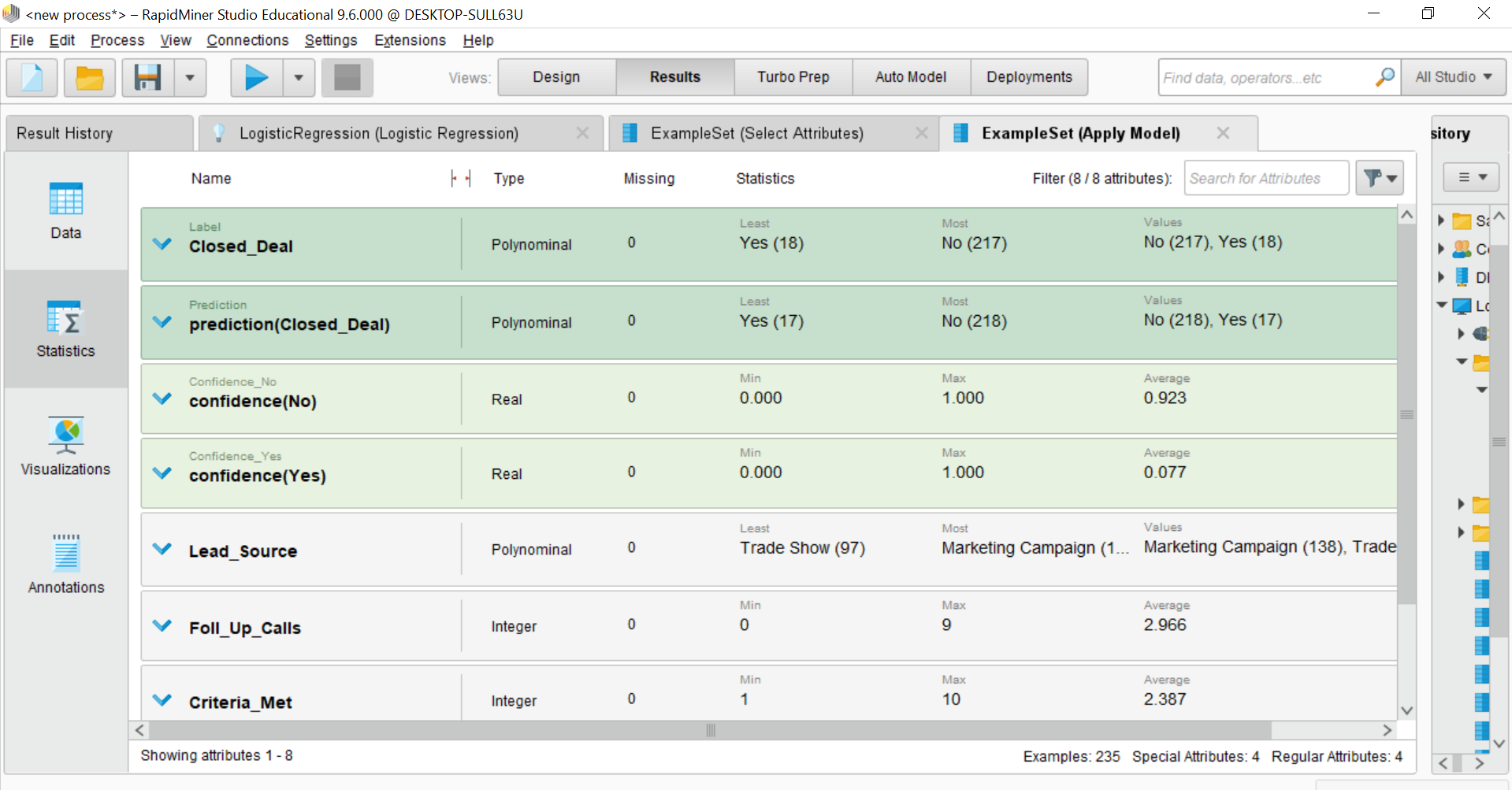
Logistic Regression



The only significant predictor variable was Criteria\_Met.



You can see by this histogram that counts the # of Criteria\_Met and uses Closed\_Deal as the color that this predictor variable separates the data very well.

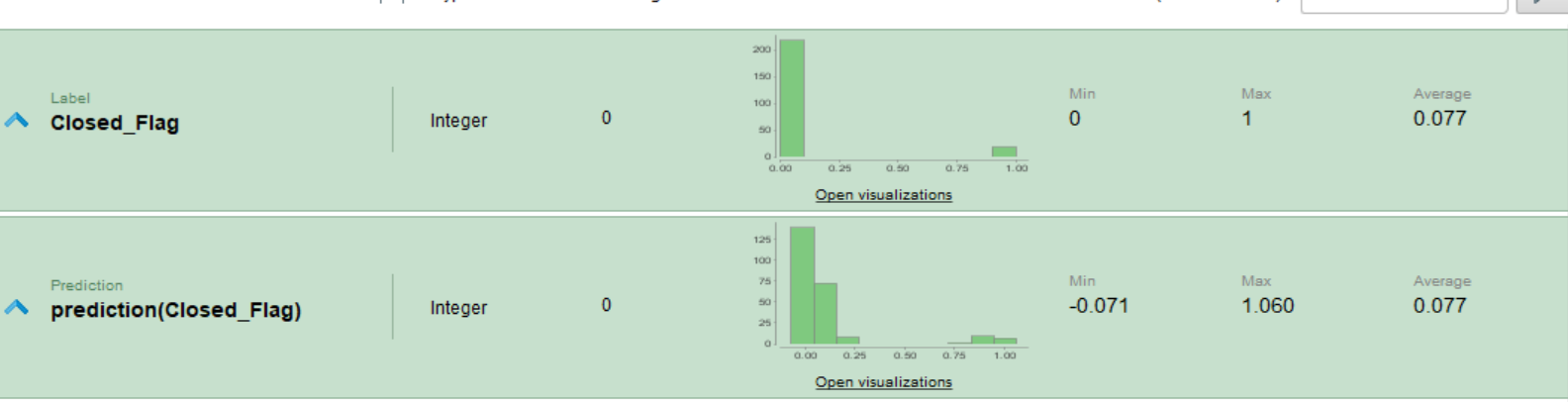


Here we see that only 1 observation was incorrectly labeled as the deal not being closed when the deal was closed.

Linear Regression:



Follow Up Calls, Criteria Met, and Sales Office are all significant predictive variables. The output of the linear regression represents the probability that the observation belongs to Closed Deal = YES. Therefore, I would interpret these coefficients as for every follow up call, the probability of the deal being closed increases .012. Similarly, for every criteria met, the probability increases .121. The Sales Office is a bit tricky since it represents a North/South location instead of a numeric property. However, I believe North = 1 so I believe when the location is the North Sales office, the probability drops .122.



If the threshold was .5, it looks like the Linear Regression would do a great job, perfect it seems like, identifying the clients that they closed.

5. Read: <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/multicollinearity-in-regression/>

Results: This article discusses multicollinearity, the repercussions of seeing significant multicollinearity in your datasets, and how to address it. Multicollinearity is a term that describes seeing correlation between predictor variables. As a result of high multicollinearity, the coefficients supplied by your model are unstable and unreliable. It could cause some significant relationships to be insignificant and high variation between coefficients assigned from different sample datasets. Multicollinearity can be measured by calculating the correlation between variables or using the variance inflation factors. You can remove multicollinearity by normalizing the variables, removing the highly correlated predictors from the model, or using PCA or partial least squares. If the predictor variables are removed, R2 won’t changes much, but the coefficients assigned will. Multicollinearity does not effect the goodness of fit or how well a prediction is made.

6. Watch Salford Systems videos Part 2 and Part 4: <https://www.salford-systems.com/resources/webinars-tutorials/how-to/how-to-interpret-model-performance-with-cost-functions>

Results: Least Squares Deviation Cost, also known as mean squared error or squared residuals, is a cost function that can be applied to machine learning algorithms. The function calculates the residuals produced by the algorithm (predicted minus actual), squares the result, and sums each up. This result is then multiplied by 1/n to normalize the cost. Its advantages include simple mathematical and statistical properties. However, one major issue that can occur with this method is how sensitive it is to outliers. Seeing as how the equation ultimately ends up as an average of the squared residual, major outliers can impact the cost significantly.

Binary Classification is a general term to describe the type of machine learning problem. The outcome is two values which can be represented by TRUE/FALSE, yes or no, positive or negative, 0 or 1, etc.). In essence, a binary classification is used to attempt to identify is some sort of condition is present in an observation. For example, is a child likely to pass a class (yes/no, pass/fail,etc). Once the target is specified, a method such as Logistic Regression can be applied, and a decision of the threshold made. For example, if h(x) (the predicted value of y that is between 0 to 1 that represents the probability of the condition being true), is set to a threshold of .6, observations with a predicted y of .6 or higher will be considered TRUE, else (less than .6) FALSE. In the example above, .6 or higher would indicate a child likely to pass, otherwise fail.