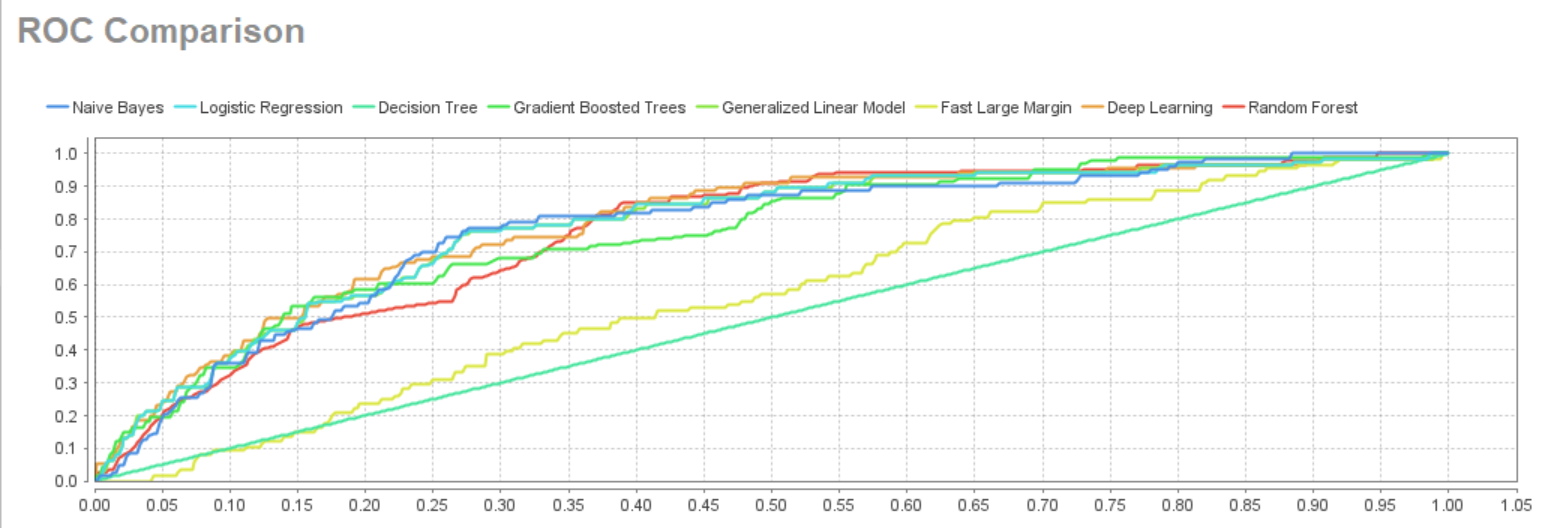
**MDS 560 Week 3 Hands-On Accelerator**

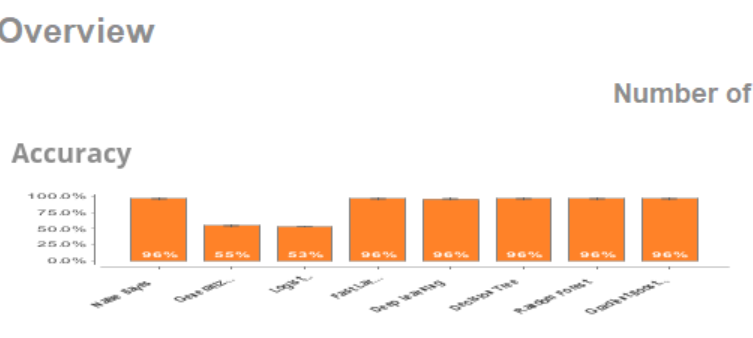
Your deliverables and hands-on activities for this week are:

1. Import dataset 3a dataset into Rapidminer, perform classification modeling using the auto model feature as appropriate to explain and predict the labeled outcomes given the selected features/independent variables. Compare the performance of the different machine learning models employed in automodel.

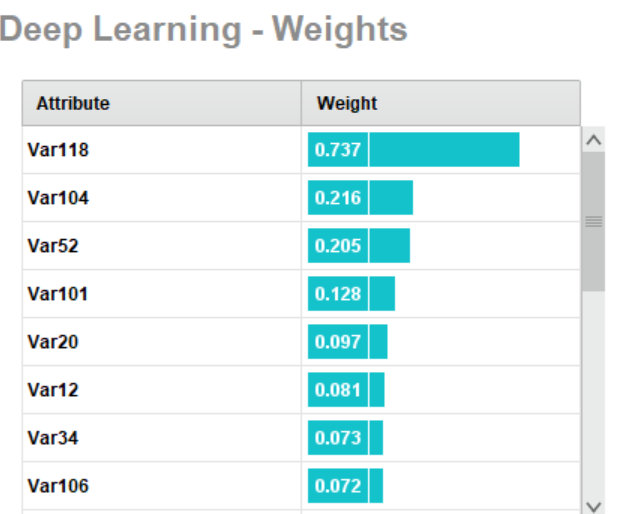
Results: The following chart compares the ROC Curve for the variety of models produced in Auto Model. You can see that Fast Large Margin and Decision Tree models performed the worse as far as maximizing the Area Under the ROC curve. The rest of the models perform similarly with Random Forest lagging behind a little bit. Deep Learning maximized the AUC overall.



You can see this very well below, but the Decision Tree and Fast Large Margin performed marginally better than Deep Learning in terms of overall accuracy. However, we know since the AUC is maximized using Deep Learning that it performs better when diving into statistics that the Performance Matrix would provide for specificity and sensitivity.

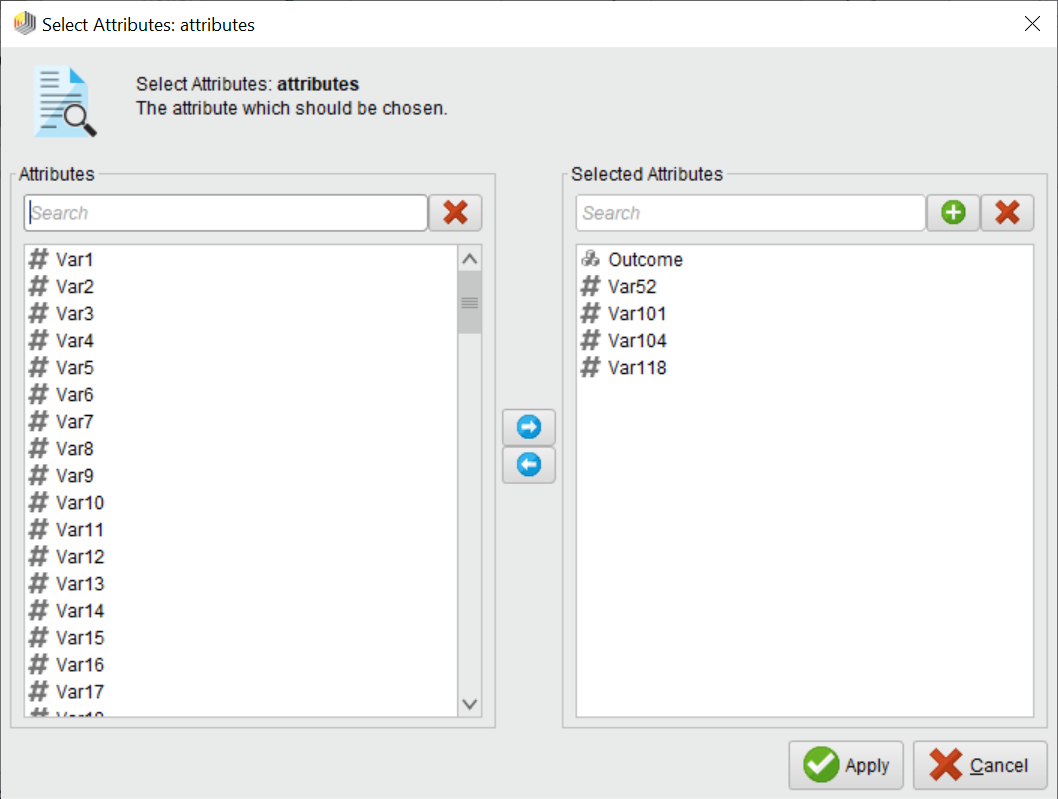


The following screenshot shows how Deep Learning assigned weights to the input features which describes feature importance. Here you see that Var118 has the highest coefficient by far with .737 which is the most indicative of a positive prediction.

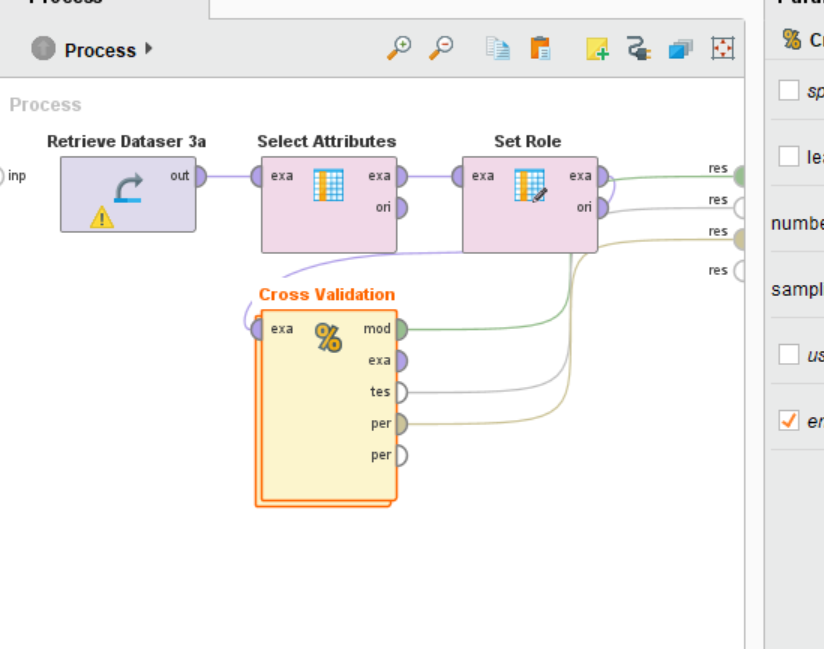


2. Using dataset 3a in Rapidminer, perform cross validation of a logistic regression model using only the most appropriate features/independent variables.

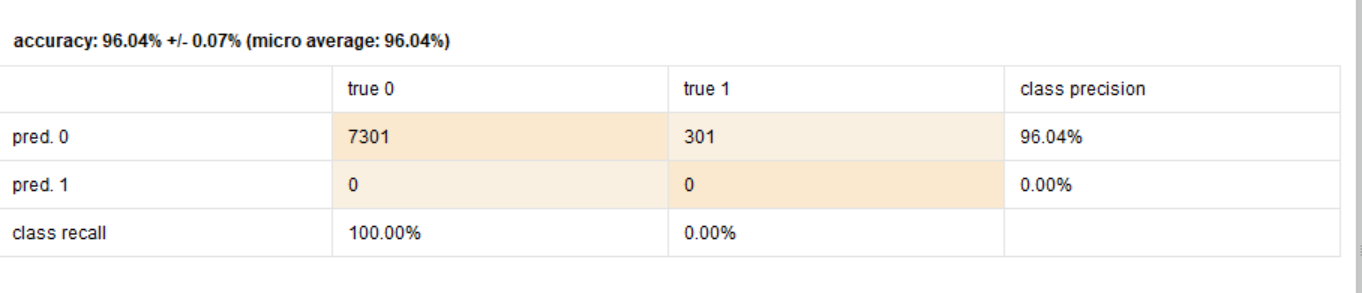
Results: In RapidMiner, I selected the top 4 variables from above to enter into the logistic regression model.

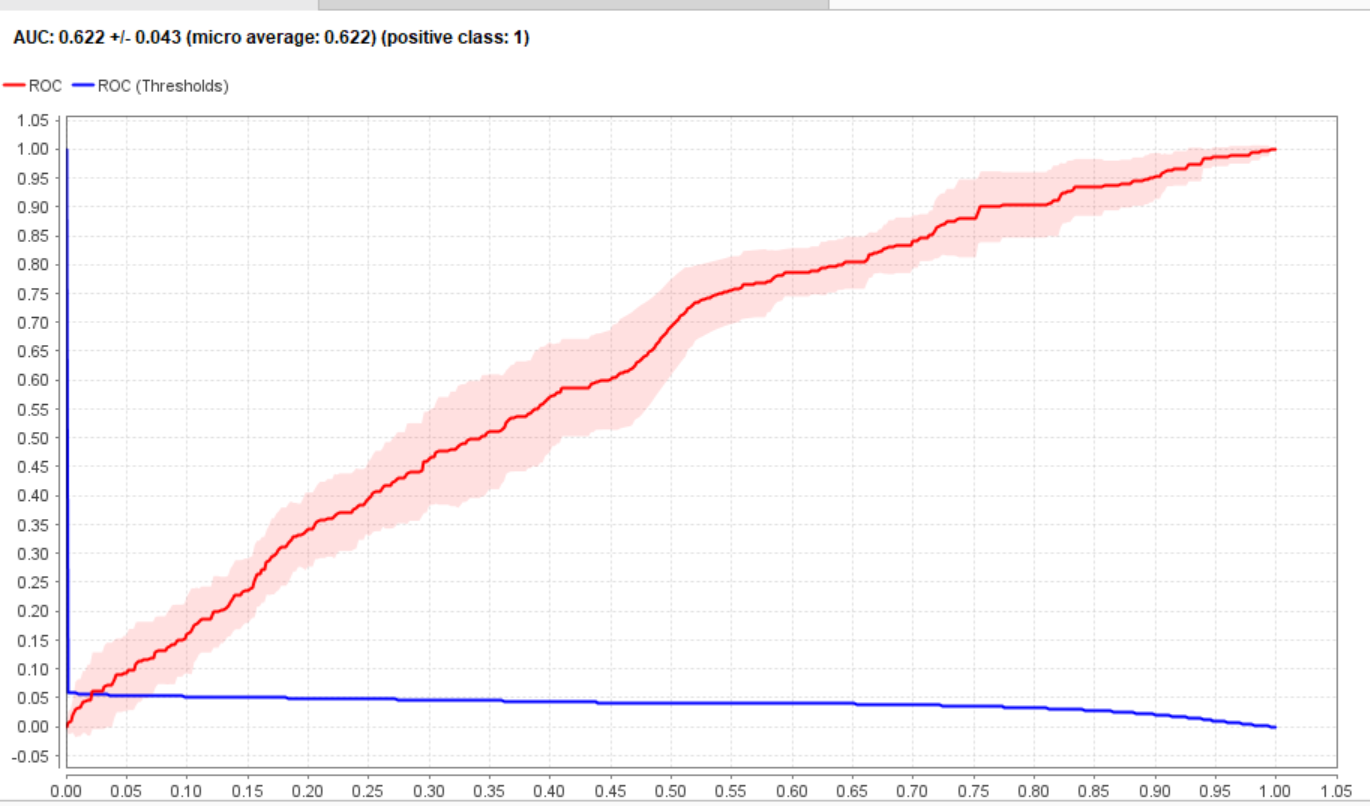


I applied a 10-fold cross validation on the dataset:



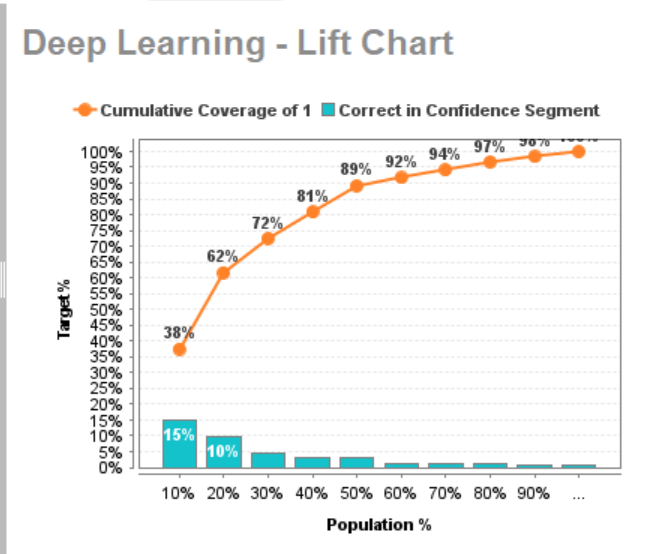
The model produced the following performance success matrix. You’ll see that it wasn’t able to predict true positives but could predict true negatives well. As a result, you will see the AUC suffer and the ROC curve to be closer to .5 which is not much better than randomly guessing true/false. However, if the overall objective was to identify true negatives, it might not be a bad fit.





3. Using the dataset 3a case, explain the usage of lift-gain charts, ROC and cross validation. Explain when validation by using hold-out data might be a better alternative to cross validation.

The lift-gain charts help you understand how much of the top x% of the dataset’s population are predicted positives. In a classification model, the predicted positives will always have the higher h(x). Therefor, if you sort by this, then the most likely positive observations will fall within the first so many percentiles of the population. The lift-gain chart helps you visualize this. The lift chart below lets you know that the first 10% of the population contains 38% of the predictive positives. If you wish to spend a cost such as running a marketing campaign, targeting to the first 10% will provide that exposure to 38% of the observations that were predicted to produce a positive result (responding to the marketing campaign).

Results

ROC Curves plot sensitivity and specificity which are metrics that evaluate true positives and true negatives. The curve plots the balance between positives and negatives by the possible threshold values. A model that predicts both positive and negative values well will maximize the AUC which provides an overall indication of how well the model performs. In contrast, if a model is better or worse at predicting positives and negatives, the curve will be flatter limiting the AUC. The higher the AUC the better the model overall performs.

Cross Validation takes a dataset and performs a test/train split multiple times (k fold times), applies a model to the train set, measure the performance on the test set, and then averages the performance results. The advantage to cross validation is being able to measure how well the model will perform on unseen data better than measuring in sample performance on the training set alone without cross validation. Cross Validation typically occurs when evaluating model performance and for hypertuning model parameters. Validating using a holdout set only is a reasonable alternative to cross validation when you have many data observations in your dataset.

4. Review the performance metrics found in the first section of the APICS SCOR documentation. Describe which KPI’s would be most relevant to your organization. Does your organization currently do a good job in measuring these indicators?

Results: I work for the Illinois Comptroller’s Office so the private supply chain environment the APICS SCOR performance metrics describes is very different from my organization. However, I think the Cost metric is similar. While a supply chain manager might look to reduce material, labor, or production costs , the Comptroller has to try to allocate funds to pay our bills according to the amount of money in Illinois’ account for the day while also keeping in mind prompt pay laws and notes that are costing us interest. The overall goal is to minimize the cost of financing for Illinois. The Comptroller tries to be as financially transparent as possible to alert the public and researchers to these hidden costs. Reports are issued monthly that summarizes prompt pay and interest charges reported by agencies. In this way, the KPI is tracked and measured.

5. Watch Salford Systems videos Parts 5, 6, 7 and 8 on Binary Classification ROC, Precision-Recall, Lift-Gains and logistic regression [*https://www.salford-systems.com/resources/webinars-tutorials/how-to/how-to-interpret-model-performance-with-cost-functions*](https://www.salford-systems.com/resources/webinars-tutorials/how-to/how-to-interpret-model-performance-with-cost-functions)

Results:

The following are the main concepts I took from watching the assigned videos.

Part 5 -

Precision and Recall - 4 events in a binary classification problem: correctly assigned positive (true positive), incorrectly assigned positive (false positive), correctly assigned negative (true negative), incorrectly assigned negative (false negative).

Precision, the ratio of true positives/(true positives + false positives) OR ration of true positives/the sum of predicted positives.

Recall (aka Sensitivity), ratio of true positives/(true positives + false negatives)

The goal is to optimize Precision and Recall meaning false positive and false negatives are close to 0. Trying to optimize only precision usually results in recall suffering and vice versa.

Specificity- true negative/(true negative + false positive)

Part 6 -

2 measurements: Recall (aka Sensitivity - focucses on positive group) & Specificity (focuses on negative group)

ROC Curve illustrates the values of recall and specificity for when the threshold ranges between the minimum value of threshold to the maximum value of threshold. Plotting these points results in a curve which helps evaluate the overall model performance. A good classifier will have a higher h(x) for true positives and lower h(x) for true negatives which creates a high area under the ROC curve. In contrast, if the classier performed poorly, the h(x) do not split the true negative and true positives as well that result in a more subtle curve that has a smaller area under the curve than the alternative.

Part 7 -

Measurements: Recall (aka Sensitivity), Support (% of population) = (true positives + false positives)/n , Base Rate (true positive+false negatives)/n = the percentage of responders (positives) in the dataset.

Gains Curve - plots sensitivity over support (essentially the % of the population. In the video's example, the study was identifying the top 20% of people likely to respond to a marketing ad. Support = .20%) Plot what the senesitivity would be amoung max and min support results in Gain Curve. Random sampling would result in more of a 45 degree angled line that splits the hyperplane into two. The angle of the gains curve is controlled by the base rate. A low base rate will and a well performing modeling will produce a curve with a lot of area underneath it while a higher base rate and a well performing model could look similar to a "worse case scenario" 45 degree angle gains curve.

Lift, plots sensetivity against (sensitivity/support).

You can convert Gains curve to ROC curve.

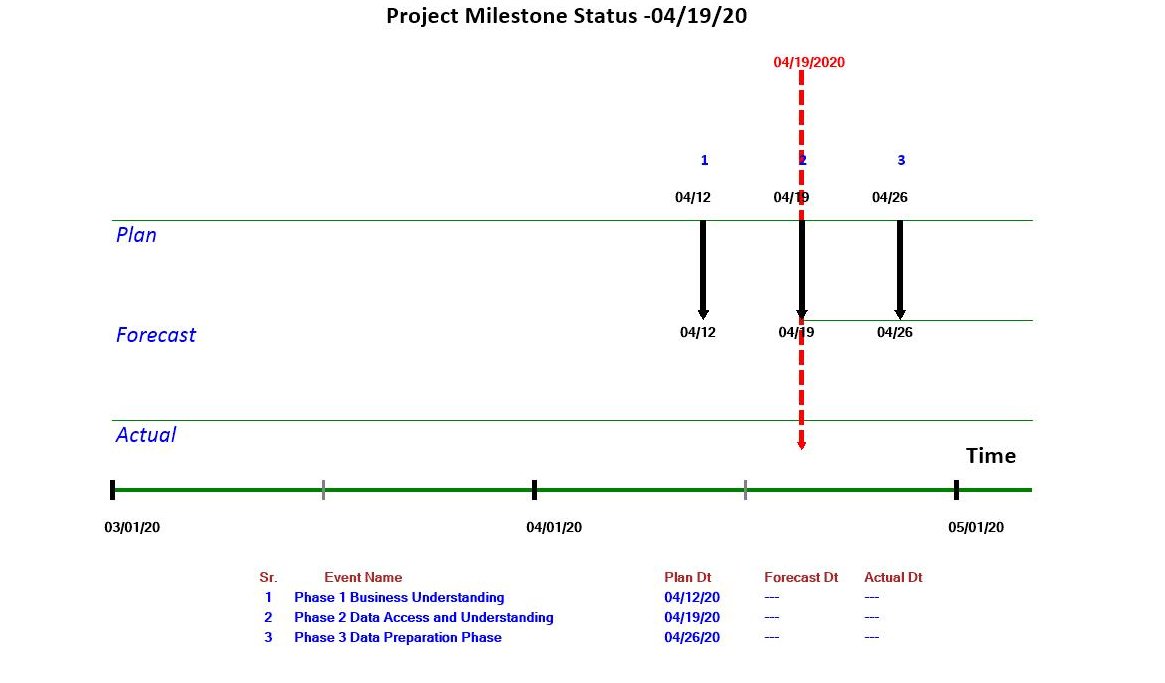
Part 8 -

Probability of a positive, constraint of being between 0 and 1.

log-odds is a way to convert from probability to a figure without that constraint. Sigmoid Funcion maps figures that is essentially betweeen 0 & 1, (from negative inifitity, getting closer to 0, to positive infinity getting closer to 1.) Log Odds and Probability can quicker be transformed into one or the other. The assumption that are allowed for interpretation with log odds can apply maximum likelihood principals.

6. For your project in PRIMMS Close out all activities for Phase 1 Business Understanding and Phase 2 Data Access and Understanding.

Results: I close out all activities for the first two phasses in PRIMMS and produced the following slip chart.



7.**Optional and Extra Credit:**

Watch the following videos from Andrew Ng (Stanford) and describe any insights that you gained regarding how the cost function for logistic regression differs from the cost function used for typical linear regression. In both cases how does the calculation of parameter values depend upon finding the minimization of the cost function?

<https://www.youtube.com/watch?v=HIQlmHxI6-0>

<https://www.bing.com/videos/search?q=logistic+cost+function&&view=detail&mid=E4E30AABEED7EAB97BB0E4E30AABEED7EAB97BB0&&FORM=VRDGAR>

<https://www.youtube.com/watch?v=blWpQyHCwHE>

<https://www.youtube.com/watch?v=W9iWNJNFzQI>

Results:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* *Optional: If needed watch the Rapidminer video describing the Auto Model functionality available in the tool.* [*https://rapidminer.com/resource/automated-machine-learning/*](https://rapidminer.com/resource/automated-machine-learning/)