MIS 3300 | Final Project

Names: **Michael D. Scovill, Michael Weston, Amata S. Tuleuova, Bryant Purnell**

Summary

This assignment is broken down into five parts: 170 pts

PART 1: Data Preparation & Data Understanding (40 points)

PART 2: Unsupervised DM (26 points)

PART 3: Supervised DM Technique 1 (28 points)

PART 4: Supervised DM Technique 2 (27 points)

PART 5: Evaluation of Models & Business Recommendations (49 points)

Each part can be started now, except Part 5. You must create and run the RapidMiner processes for Parts 3 and 4 in order to proceed with Part 5.

PART 1: Data Preparation & Data Understanding (40 points)

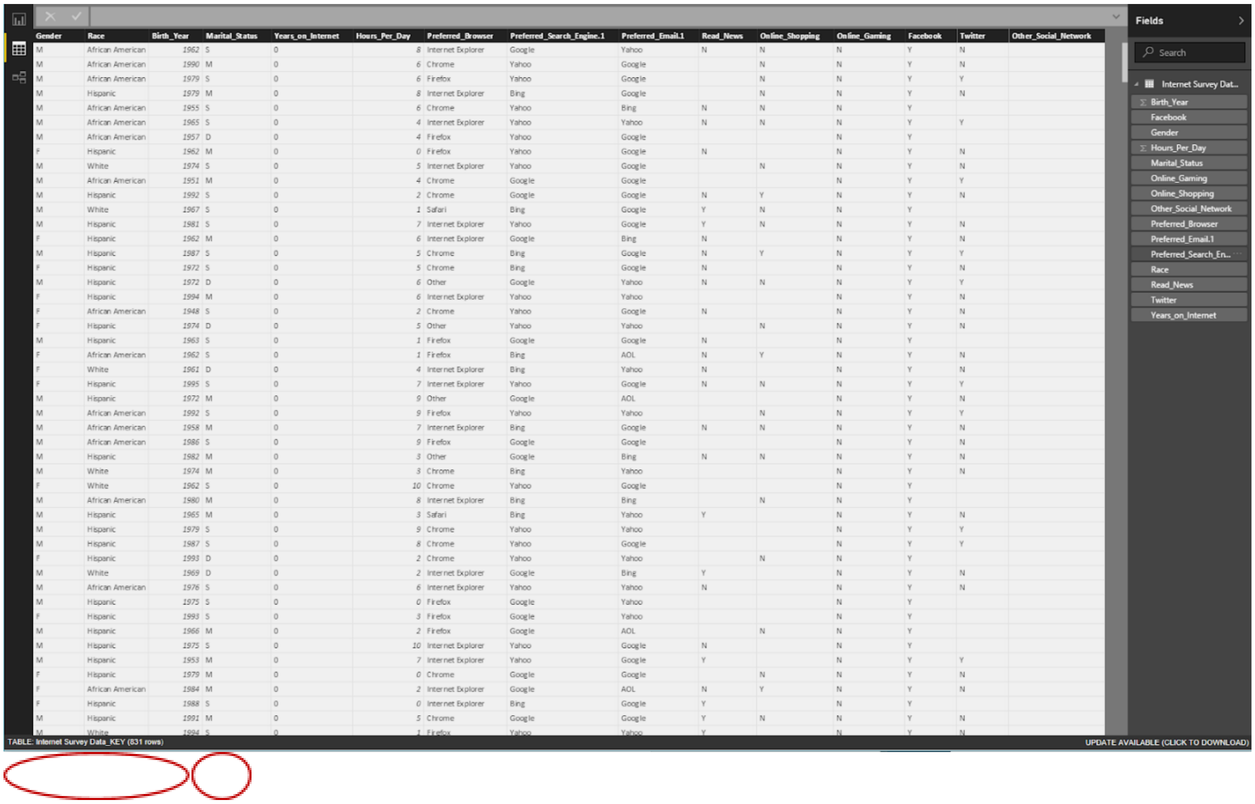
A. Data Preparation (16 pts)

Import the *manager_performance_v3.csv* dataset into Power BI, and clean/transform this data set. If you need some reminders about how to do this, revisit the data preparation module!

- Think about any ethical concerns regarding this dataset. Remove any columns that personally identify employees or could be used to discriminate against employees (sex, marital status, age, sexual orientation, etc.).
- Go through *each attribute column* and perform various data transformations necessary to cleanse the dataset. For each attribute/column, report *each data cleansing step performed and the underlying assumption* as to why the data cleansing action was performed.
    - Do not simply state that "all columns were trimmed" or restate the cleansing action itself.
    - State the *assumption* (e.g., "M" was changed to "Male" because it was assumed that "M" indicated "Male" in this dataset.).
    - Also, if no data transformations were made, state your assumption here as well (all data were assumed to be correct/clean).
- Pay attention to column formats *after* you're finished editing each column.

When you're finished performing data transformations in Power BI, take a screenshot of the Query Editor Window (this is a back-up), then click close and apply. Ensure you're on the Data Tab, then take another screenshot here of the data tab. The screenshot does not need to illustrate all rows, but do include the total number of rows shown at the bottom of the data tab table and all remaining attribute columns. Include this screenshot of the data tab (like the example below).

Example of .pbix screenshot (data tab circled to the left; number of rows circled at the bottom; and all attributes showing in the table):



Save your work as a .pbix file in case you need it later or would like to create data visualizations in Power BI (also later).

For this portion of the assignment, add the list of assumptions and a screenshot of your .pbix file data tab illustrating the cleansed dataset. Do not submit your .pbix file. The group should continue adding the below Final Project parts/requirements to this document.

Data Transformations & Assumptions

Manager ID: all data were assumed to be clean.

First Name and Last Name: both were removed because it was assumed that they could personally identify employees.

Age: this column was removed because it was assumed that it could be used to discriminate against employees.

Time Employed: “O” was changed to 0 because it was assumed that “O” indicated 0 in this dataset. The column type was changed from “Text” to the “Whole number” because it was assumed that the number of years is not text, and this would prevent the numbers from being sorted in a specific order.

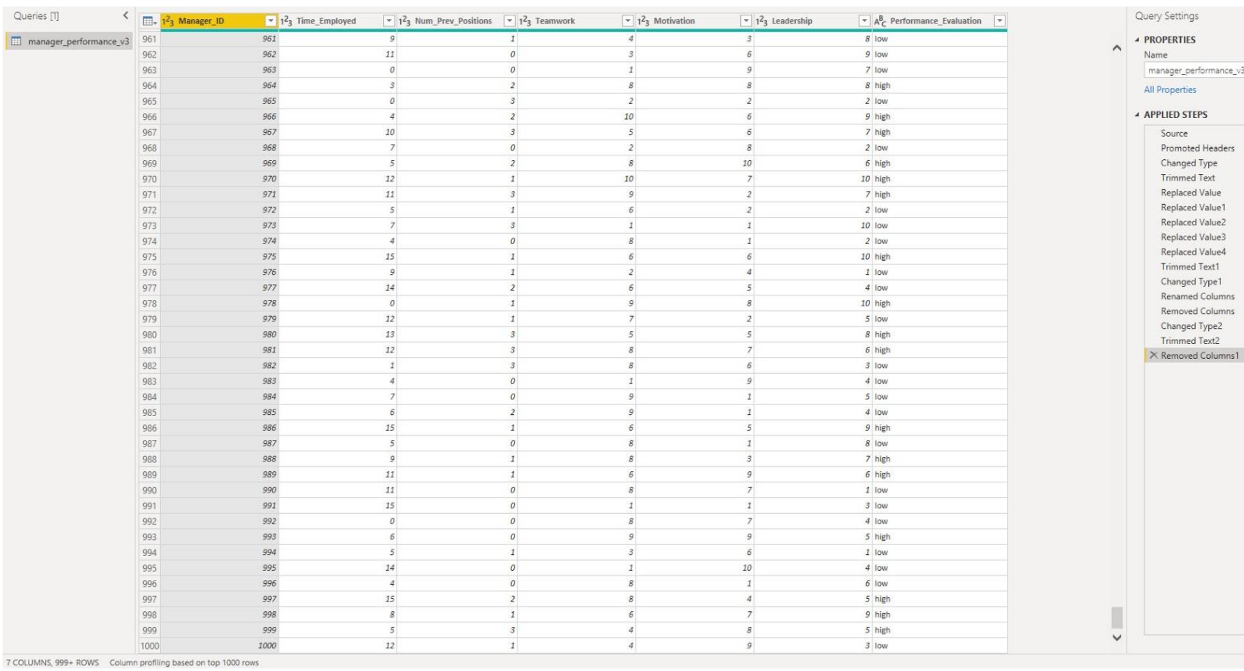
Num Previous Positions: “zero” was changed to 0 because it was assumed that “zero” indicated 0 in this dataset. The column type was changed from “Text” to the “Whole number” because it was assumed that the number of previous management positions would include whole numbers.

Teamwork: The column type was changed from “Decimal Number” to the “Whole number” because it was assumed that the peer rating would include whole numbers 1-10.

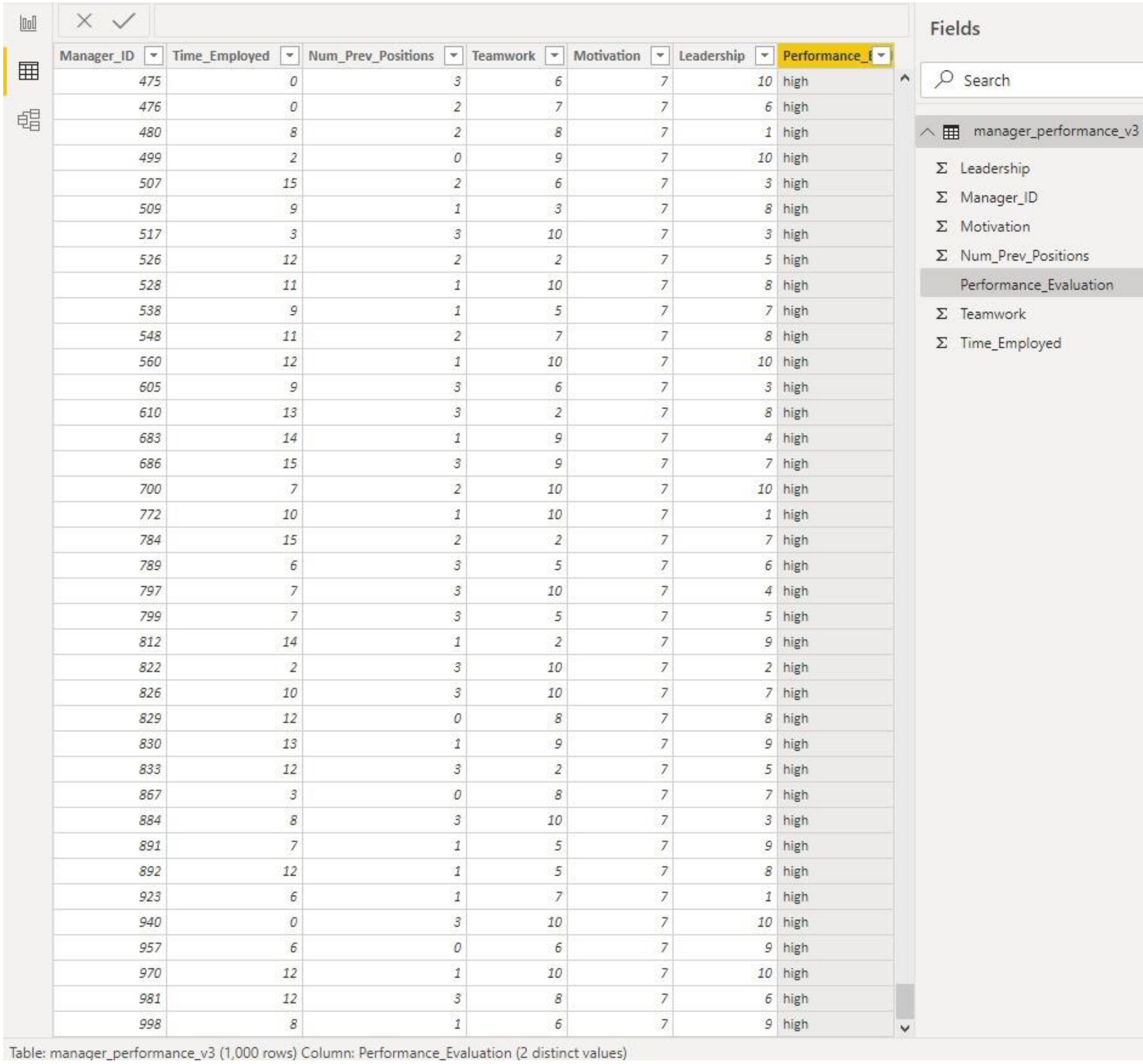
Motivation and Leadership: all data were assumed to be correct.

Performance Evaluation: “highgh” and “loww” were changed to “high” and “low” because it was assumed that “highgh” and “loww” indicated “high” and “low” in this dataset.

Screenshot



	Manager_ID	Time_Employed	Num_Prev_Positions	Teamwork	Motivation	Leadership	Performance_Evaluation
961	961	9	1	4	3	8	low
962	962	11	0	3	6	9	low
963	963	0	0	1	9	7	low
964	964	3	2	8	8	8	high
965	965	0	3	2	2	2	low
966	966	4	2	10	6	9	high
967	967	10	3	5	6	7	high
968	968	7	0	2	8	2	low
969	969	5	2	8	10	6	high
970	970	12	1	10	7	10	high
971	971	11	3	9	2	7	high
972	972	5	1	6	2	2	low
973	973	7	3	1	1	10	low
974	974	4	0	8	1	2	low
975	975	15	1	6	6	10	high
976	976	9	1	2	4	1	low
977	977	14	2	6	5	4	low
978	978	0	1	9	8	10	high
979	979	12	1	7	2	5	low
980	980	13	3	5	5	8	high
981	981	12	3	8	7	6	high
982	982	1	3	8	6	3	low
983	983	4	0	1	9	4	low
984	984	7	0	9	1	5	low
985	985	6	2	9	1	4	low
986	986	15	1	6	5	9	high
987	987	5	0	8	1	8	low
988	988	9	1	8	3	7	high
989	989	11	1	6	9	6	high
990	990	11	0	8	7	1	low
991	991	15	0	1	1	3	low
992	992	0	0	8	7	4	low
993	993	6	0	9	9	5	high
994	994	5	1	3	6	1	low
995	995	14	0	1	10	4	low
996	996	4	0	8	1	6	low
997	997	15	2	8	4	5	high
998	998	8	1	6	7	9	high
999	999	5	3	4	8	5	high
1000	1000	12	1	4	9	3	low



Manager_ID	Time_Employed	Num_Prev_Positions	Teamwork	Motivation	Leadership	Performance_E
475	0	3	6	7	10	high
476	0	2	7	7	6	high
480	8	2	8	7	1	high
499	2	0	9	7	10	high
507	15	2	6	7	3	high
509	9	1	3	7	8	high
517	3	3	10	7	3	high
526	12	2	2	7	5	high
528	11	1	10	7	8	high
538	9	1	5	7	7	high
548	11	2	7	7	8	high
560	12	1	10	7	10	high
605	9	3	6	7	3	high
610	13	3	2	7	8	high
683	14	1	9	7	4	high
686	15	3	9	7	7	high
700	7	2	10	7	10	high
772	10	1	10	7	1	high
784	15	2	2	7	7	high
789	6	3	5	7	6	high
797	7	3	10	7	4	high
799	7	3	5	7	5	high
812	14	1	2	7	9	high
822	2	3	10	7	2	high
826	10	3	10	7	7	high
829	12	0	8	7	8	high
830	13	1	9	7	9	high
833	12	3	2	7	5	high
867	3	0	8	7	7	high
884	8	3	10	7	3	high
891	7	1	5	7	9	high
892	12	1	5	7	8	high
923	6	1	7	7	1	high
940	0	3	10	7	10	high
957	6	0	6	7	9	high
970	12	1	10	7	10	high
981	12	3	8	7	6	high
998	8	1	6	7	9	high

B. Data Understanding (24 pts):

Return to Canvas and download the *manager_performance_v3_clean.csv*. This file is provided so that any errors potentially made during data cleansing do not result in subsequent errors/deductions for the remaining portions of the assignment.

Thoroughly explore the data by creating and interpreting descriptive statistics. Specifically:

1. Create a descriptive statistics table below that includes the sample size, mean, median, min, max, and standard deviation for all continuous variables in the dataset. This table should conform with the standard format provided in the textbook and Data Understanding Guide. Do not attach an Excel file.

Variable	n	x	M	min.	max	s
time_Employed	1000	7.468	8	0	15	4.597
num_Prev_Positions	1000	1.481	1	0	3	1.126

2. From your descriptive statistics table, create at least two hypotheses regarding factors driving employee performance evaluations. These hypotheses should not be simple restatements of the facts shown in the tables, but instead should reflect your thoughts about the potential underlying causes of these results (i.e., what might be causing the results seen in the table?). Explain any rationale behind your hypotheses as needed to clarify your line of thinking.

The amount of time someone has been employed in a managerial position could indicate that they will have a higher performance evaluation score.

If someone has had many previous positions as a manager that could mean that they are not performing well as a manager.

3. Create a correlation matrix in RapidMiner including the dependent variable and all candidate independent variables (any variable that might predict the value of the DV). Paste a screenshot of the matrix here.

Attributes	Perfor...	Time_E...	Num_Pr...	Teamw...	Motivati...	Leader...
Performance_Evaluation = low	1	-0.222	-0.311	-0.369	-0.383	-0.325
Time_Employed	-0.222	1	0.062	-0.030	0.051	-0.042
Num_Prev_Positions	-0.311	0.062	1	-0.012	0.024	-0.010
Teamwork	-0.369	-0.030	-0.012	1	-0.009	-0.006
Motivation	-0.383	0.051	0.024	-0.009	1	-0.021
Leadership	-0.325	-0.042	-0.010	-0.006	-0.021	1

4. Are there any variable pairs that are multicollinear (use correlation coefficient value of 0.6)? Explain your answer.

No, there is no correlation between variable pairs because there’s no value at or above 0.6.

5. Look through this data set and, ignoring the ID, identify the types of data variables therein (nominal, ordinal, ratio, or interval).

Time Employed: ratio
Num Prev Positions: ratio
Teamwork: Ordinal
Motivation: Ordinal

Leadership: Ordinal
Performance Evaluation: Ordinal

PART 2: Unsupervised DM (26 points)

Download the *manager_performance_v3_clean.csv dataset* from Canvas. This file is provided so that any errors potentially made during data cleansing do not result in subsequent errors/deductions for the remaining portions of the assignment.

In RapidMiner, import the clean dataset and conduct an unsupervised data mining technique appropriate for this dataset. Think about the data variable types in this dataset and the business question, then choose from association rules analysis or clustering analysis.

HINT: Revisit the 'summary slide' for each of the unsupervised data mining techniques to remember what type(s) of data variables can be input into these models.

Choose an appropriate and informative unsupervised data mining model operator (we've used it before in class). Think about what we're interested in figuring out with this manager performance dataset (performance evaluation) and how many classes we have for this attribute. Change one parameter based on this. Uncheck 'determine good start values'.

Screenshots will be included in Part B below.

A. Identify which model operator was selected *and why*. The *why* should focus on the types of data variables in the manager performance data set. Also discuss any parameters changed and *why*.

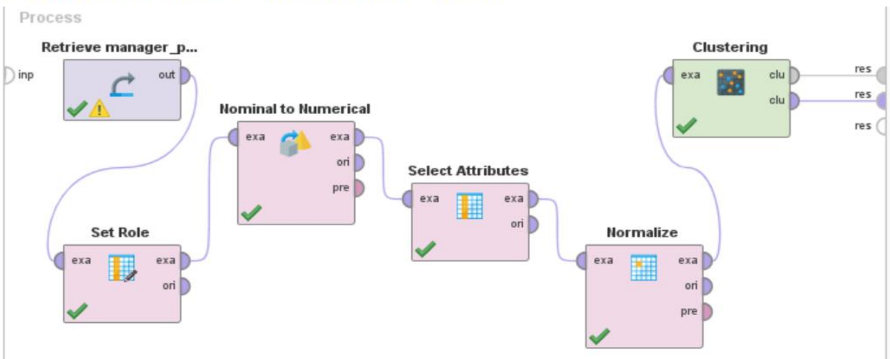
The model operator that was selected was **Cluster Analysis** because for the unsupervised data mining analysis, we have data that is both quantitative e.g. Time_Employed, Num_Prev_Positions, Teamwork, Motivation, and Leadership. We also have qualitative data e.g. Performance_Evaluation (which will need to be changed to numerical). Which is all acceptable data for this analysis. We are also looking to place some of the good managers into natural groups to help us determine which managers would be best, which is why cluster analysis is a good fit for this analysis.

Some of the parameter changes that will be made are Performance Evaluation will need to be converted from Nominal to Numerical. Manager will need to be removed through the "Set Role" or the "Select Attributes" operators. To not overweight the analysis we have removed Performance_Evaluation = low. And we also used the normalize operator to ensure that none of the data outweighs any other data.

B. Include screenshots of your RapidMiner process window and relevant results screens and interpret these results. Revisit the previous exercises for the chosen model to remember what the relevant

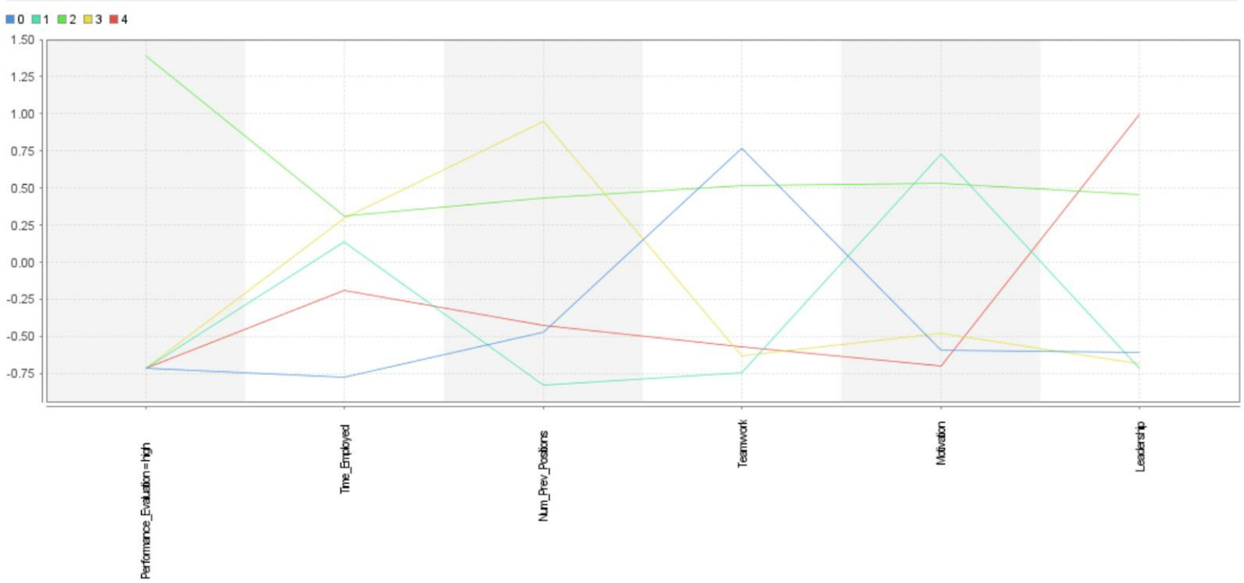
Cluster Model

```
Cluster 0: 178 items
Cluster 1: 159 items
Cluster 2: 341 items
Cluster 3: 150 items
Cluster 4: 172 items
Total number of items: 1000
```



results screens and interpretations should focus on.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Performance_Evaluation = hi...	-0.719	-0.719	1.389	-0.719	-0.719
Time_Employed	-0.776	0.139	0.309	0.294	-0.194
Num_Prev_Positions	-0.472	-0.829	0.432	0.946	-0.427
Teamwork	0.765	-0.742	0.513	-0.632	-0.570
Motivation	-0.594	0.729	0.533	-0.478	-0.698
Leadership	-0.608	-0.713	0.452	-0.688	0.993



From this data, we created 5 clusters. We can see that there does not seem to be any obvious outliers from our Cluster Analysis because each of the clusters have similar amounts of items (Found on the Cluster Model: 178, 159, 341, 150, and 172.)

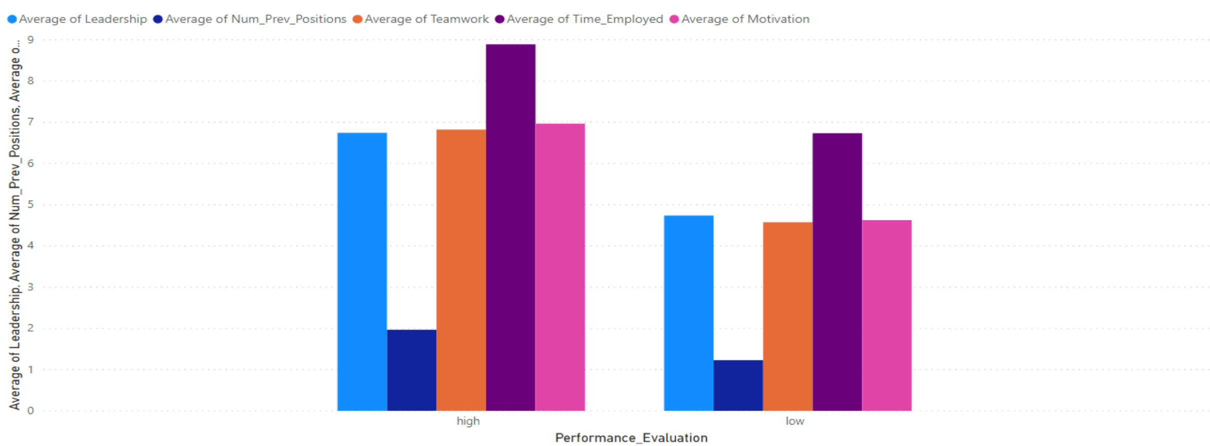
While analysing the data, we want to know which of these clusters will most likely have the highest performance evaluation. **Cluster 2** is the only cluster in our analysis that has a positive Performance Evaluation score (Z-Score of 1.389). This means that the participants in Cluster 2 seem to have higher performance evaluations which on average is 1.389 standard deviations above the mean.

In Cluster 2 we see that the Independent Variables are all about 0.25 and 0.50 standard deviations above the mean for each attribute (Time_Employed = 0.309, Num_Prev_Positions 0.432, Teamwork 0.513, Motivation 0.533, Leadership 0.452). While the other clusters have Performance_Evaluation = high scores less than -0.700 (cluster 0, 1, 3, and 4 all have a Performance_Evaluation = high of -0.719), which means they are all about 0.719 standard deviations below the mean.

Cluster 2 will have the managers the HR director will want for the open position because on average they have higher Performance_Evaluation = High scores compared to the other clusters. While reviewing the independent variables for cluster 2, we see that there is not one attribute that accounts for a good manager, but instead all of the independent variables have a score of 0.25 to 0.50 standard deviations above the mean. For example, Cluster 0 might score high with regards to Teamwork but scores low in almost every other category. This is very similar with Clusters 0, 1, 3, and 4.

From this cluster analysis we see that The HR Director should not look for someone that scores high on just one of these Independent variables, but instead should look for someone who scores consistently higher than average on all of the variables. **High performing managers seem to be balanced.**

C. Create at least one visualization (in RapidMiner or Power BI) with a caption or description about how this visualization contributes towards the *meaningful* interpretation of the manager performance data. You cannot use any visualizations automatically generated by RapidMiner. You must draw from the visualization portion of this course and create your own relevant visualization, label it, and include a brief caption.



Average Independent Variable values by Performance Evaluation :
This graph shows that the managers with High Performance Evaluations score higher then the Low Performance Evaluations in every variable on average as seen in the graph above. Which shows up that high performing managers are not determined by high scores in one variable, but instead they are balanced and score high in each Variable.

PART 3: Supervised DM Technique 1 (28 points)

Using the *manager_performance_clean_v3.csv* file from Canvas, conduct a supervised data mining technique in RapidMiner. You will compare this chosen supervised data mining technique to a different supervised data mining technique later (Part 5). Think about the data variables and DV in this dataset, then choose from linear regression or decision tree analysis.

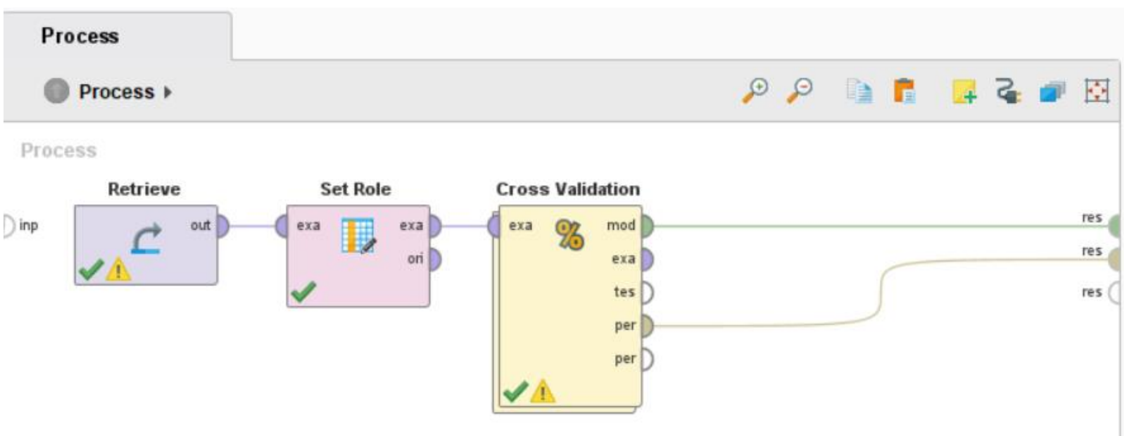
Use a cross-validation operator and create a nested process with your model operator, Apply Model, and the correct type of Performance operator. For the model operator, set minimal leaf size to 10 and ensure the maximal depth is set to 10. For the Performance operator, select: accuracy, classification error, kappa, lift, and f-measure.

A. Identify which model operator was chosen *and why*. Focus on the data variable types and the DV in the manager performance data set. HINT: Think about different types of supervised models we've learned about – regression versus classifiers.

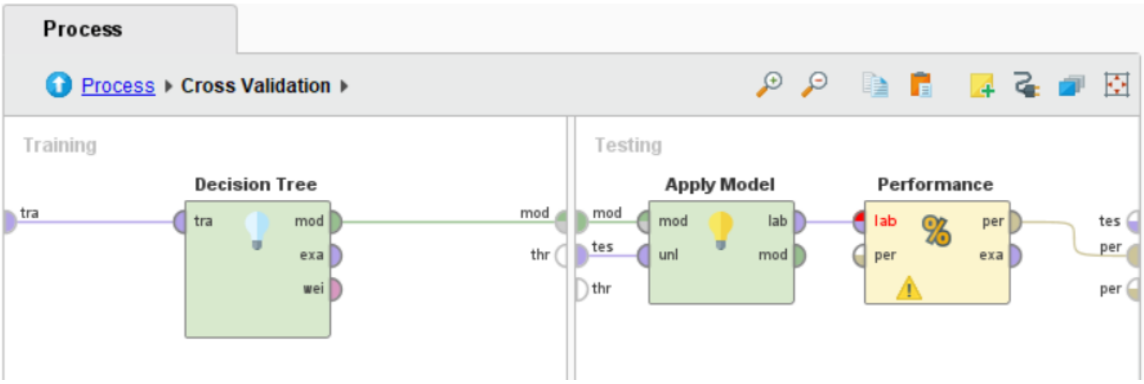
We chose to use the Decisions Trees model to predict the value of the categorical DV, which is Performance_evaluation, given the input of all the other independent variables.

B. Include screenshots of your processes and relevant results screens and *interpret* these results based on the positive class being performance_evaluation = low. Do not just restate results. Tell me what these results mean. Revisit the previous exercise for the chosen model to remember what the relevant results screens and interpretations should focus on.

Main Process Window:



Nested Process Window (within the cross-validation operator):



The tree below helps us interpret the model we created by showing us a prediction based on the input of the independent variables. It starts us off with the Motivation independent variable, then based on the value of that variable, it will direct us to the next independent variable where we look at that value, and then continuing down the tree until we see what the predicted value for performance_evaluation is based on the inputs of the IV's.

The below Confusion matrix tells us the overall accuracy of the model, which is 94.10%, as well as the actual results and predictions of the model. In the first row, the model predicted that an observation of pred. high was accurate 320 times and inaccurate 38 times, which leads to an accuracy or class precision of 89.39%. On the next row, we see that the model predicted that a result of pred. low was accurate 621 times and inaccurate 21, which equals a 96.73% accuracy rate. These accuracy rates tell us that our model did much better in predicting low than it did predicting high.

accuracy: 94.10% +/- 2.28% (micro average: 94.10%)

	true high	true low	class precision
pred. high	320	38	89.39%
pred. low	21	621	96.73%
class recall	93.84%	94.23%	

The description tab within the Performance Vector tells us the results of the model in its entirety, with the accuracy being 94.10%, the Classification error being 5.90%, the Kappa being 0.870 , the Lift being 1.4688, and the F-measure being 0.9546.

The classification error measures how many errors we had in our matrix, and compares it to all of our observations. It ranges from 0-1, and lower is better. With the error rate being 5.90% we have a very low error rate which means our model is performing very well.

The Kappa value measures how well our model is performing the above change. It ranges from 0-1, the closer to 1 the better. This kappa is performing above chance, so the value is right where we want it.

Lift ranges from 0 to infinity, with higher being better. This model performs 146.88 times better than a naïve prediction at identifying truly positive cases.

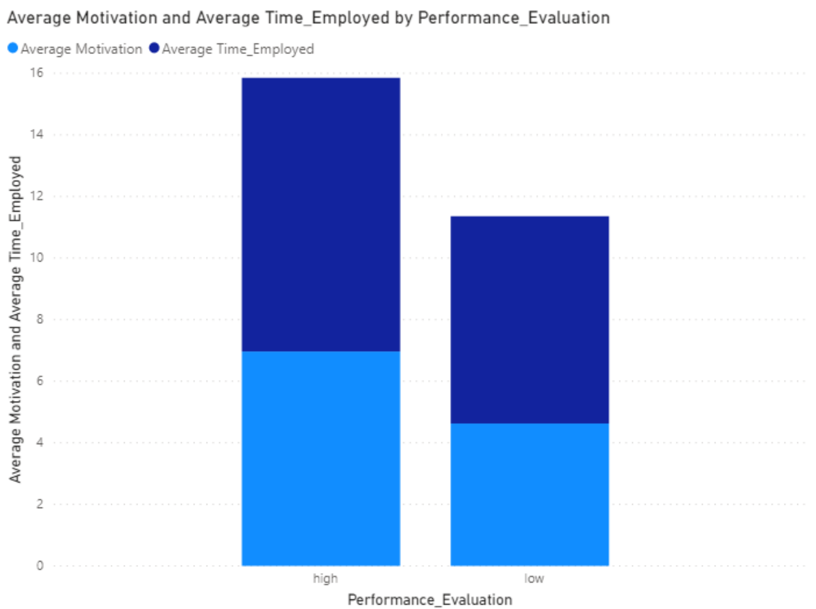
F-measure ranges from 0-1, and closer to 1 is better. The f-measure is 0.9546 which is in the range we want it to be.

PerformanceVector

```
PerformanceVector:
accuracy: 94.10% +/- 2.28% (micro average: 94.10%)
ConfusionMatrix:
True:   high    low
high:   320     38
low:    21      621
classification_error: 5.90% +/- 2.28% (micro average: 5.90%)
ConfusionMatrix:
True:   high    low
high:   320     38
low:    21      621
kappa: 0.870 +/- 0.050 (micro average: 0.870)
ConfusionMatrix:
True:   high    low
high:   320     38
low:    21      621
lift: 146.88% +/- 3.75% (micro average: 146.78%) (positive class: low)
ConfusionMatrix:
True:   high    low
high:   320     38
low:    21      621
f_measure: 95.46% +/- 1.76% (micro average: 95.47%) (positive class: low)
ConfusionMatrix:
True:   high    low
high:   320     38
low:    21      621
```

C. Create at least one visualization (in RapidMiner or Power BI) with a caption or description about how this visualization contributes towards the *meaningful* interpretation of the manager performance data. You cannot use a results screens automatically generated by RapidMiner. You must draw from the visualization portion of this course and create your own relevant visualization, label it, and include a brief caption.

The below table shows the Performance Evaluations compared to the Average Time_Employed, and Average Motivation. I used the variables in my visualization because the value of the variables heavily impact the evaluation decision.



PART 4: Supervised DM Technique 2 (27 points)

Using the *manager_performance_clean_v3.csv*, conduct another type of supervised data mining technique in RapidMiner. It cannot be the same modeling type (decision tree or linear regression) used in Part 3 with different parameters. But it should be type of data mining technique that can be used on this data set, so again think about the data variable types and the DV.

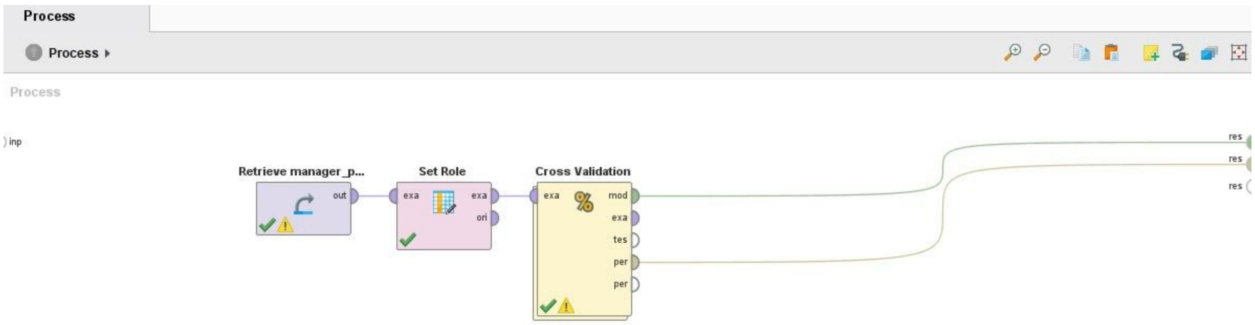
Use the cross-validation operator and set up the nested process with the model operator, apply model operator, and performance operator. For the model operator, uncheck 'remove collinear columns'. Be sure to choose the correct type of Performance operator, given the DV in this data set. In the Performance operator, check accuracy, classification error, kappa, lift, and f-measure.

A. Identify which model operator was chosen *and why*. Focus on the data variable types and the DV in the manager performance data set. HINT: Think about different types of supervised models we've learned about – regression versus classifiers – and don't reuse the model type used in Part 3.

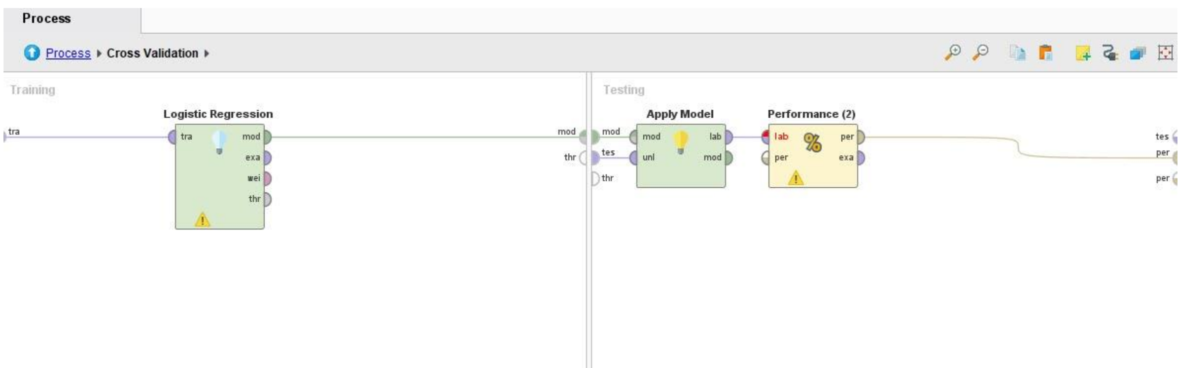
We chose logistic regression to predict the probability of the outcome of the performance given the input of the other variables.

B. Include screenshots of your processes and relevant results screens and *interpret* these results. Do not just restate results. Tell me what these results mean. Revisit the previous exercise for the chosen model to remember what the relevant results screens and interpretations should focus on.

The main process window:



The nested process window (within the cross-validation operator):



The table that helps to interpret Logistic Regression that includes important columns such as the actual coefficient that we would plug into our logistic regression equation, the standardized coefficient that we would use to determine which IV is having the greatest impact on the log odds of DV, p-value that we would use to determine a significant impact or not of our IV.

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Time_Employed	-0.294	-1.352	0.032	-9.256	0
Num_Prev_Positions	-1.749	-1.970	0.157	-11.173	0
Teamwork	-0.781	-2.257	0.063	-12.332	0
Motivation	-0.797	-2.308	0.065	-12.314	0
Leadership	-0.744	-2.177	0.062	-11.998	0
Intercept	19.400	2.087	1.359	14.275	0

The description tab within the logistic regression results tab indicates the model performance metrics such as R^2 is 0.67, which indicates this model explains 66.9% of the variance in the predicted logit value and AUC is 0.96, which means this model is performing better than random chance (50%).

Logistic Regression Model

```
Model Metrics Type: BinomialGLM
 Description: N/A
 model id: rm-h2o-model-logistic_regression-732132
 frame id: rm-h2o-frame-logistic_regression-825977
 MSE: 0.07441281
 RMSE: 0.2727871
 R^2: 0.6688629
 AUC: 0.9622684
 pr_auc: 0.9805894
 logloss: 0.2334271
 mean_per_class_error: 0.122673206
 default threshold: 0.4462442696094513
 CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
        high  low  Error         Rate
  high   277   64  0.1877     64 / 341
   low    38  621  0.0577     38 / 659
Totals   315  685  0.1020  102 / 1,000
Gains/Lift Table (Avg response rate: 65.90 %, avg score: 40.15 %):
```

The Confusion Matrix shows the overall model accuracy, which is 89%, actual results in the columns and the predictions in the rows. These both tell us that in the first row the model predicted an observation of pred. high was accurate 282 times and inaccurate 51 times. This totals out to an 84.68% accuracy rate or class precision. On the next row, we see that the model predicted a result of pred. low was accurate 608 times and inaccurate 59. This equates to a 91.15% accuracy rate. An 84.68% accuracy rate when predicting pred. high and a 91.15% accuracy rate when predicting pred. low tells us that our model did a better job predicting the pred. low than it did the pred. high.

accuracy: 89.00% +/- 1.76% (micro average: 89.00%)

	true high	true low	class precision
pred. high	282	51	84.68%
pred. low	59	608	91.15%
class recall	82.70%	92.26%	

The description tab within the performance Vector results tab indicates the metrics such as repeats of the confusion matrix, the accuracy (89%), Kappa (0.754), Lift (138.47), and F-measure (0.9171).

Kappa measures how well our model is performing the above change. It ranges from 0-1, the closer to 1 the better. This kappa is performing above chance, so the value is good.

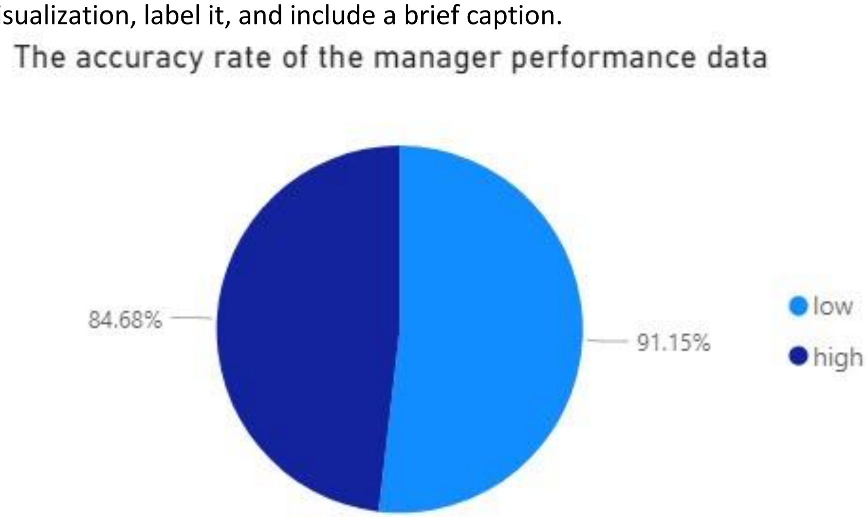
Lift ranges from 0 to infinity and the higher the better. This model performs 138 times better than a naïve prediction at identifying truly positive cases (the predicted low performance of managers).

F-measure ranges from 0-1, closer to 1 is better. Therefore, this f-measure is very good.

PerformanceVector

```
PerformanceVector:
accuracy: 89.00% +/- 1.76% (micro average: 89.00%)
ConfusionMatrix:
True:   high    low
high:   282     51
low:    59      608
classification_error: 11.00% +/- 1.76% (micro average: 11.00%)
ConfusionMatrix:
True:   high    low
high:   282     51
low:    59      608
kappa: 0.754 +/- 0.040 (micro average: 0.754)
ConfusionMatrix:
True:   high    low
high:   282     51
low:    59      608
lift: 138.47% +/- 3.98% (micro average: 138.32%) (positive class: low)
ConfusionMatrix:
True:   high    low
high:   282     51
low:    59      608
f_measure: 91.71% +/- 1.32% (micro average: 91.70%) (positive class: low)
ConfusionMatrix:
True:   high    low
high:   282     51
low:    59      608
```

C. Create at least one visualization (in RapidMiner or Power BI) with a caption or description about how this visualization contributes towards the *meaningful* interpretation of the manager performance data. You cannot use a results screen automatically generated by RapidMiner. You must draw from the visualization portion of this course and create your own relevant visualization, label it, and include a brief caption.



The chart shows the class precision metrics, which includes the insignificant difference between low and high performance among all managers.

PART 5 – Evaluation of Models & Business Recommendations (49 points)

Compare the two supervised data mining models from Parts 3 and 4 above.

A. Recopy each Performance Vector results tab (description tab & performance tab/confusion matrix) outputs from the two supervised models above (Parts 3B & 4B).

<table>
<tr><td>Logistic Regression</td><td>Decision Tree</td></tr>
<tr><td>PerformanceVector

PerformanceVector:
accuracy: 89.00% +/- 1.76% (micro average: 89.00%)
ConfusionMatrix:
True: high low
high: 282 51
low: 59 608
classification_error: 11.00% +/- 1.76% (micro average: 11.00%)
ConfusionMatrix:
True: high low
high: 282 51
low: 59 608
kappa: 0.754 +/- 0.040 (micro average: 0.754)
ConfusionMatrix:
True: high low
high: 282 51
low: 59 608
lift: 138.47% +/- 3.98% (micro average: 138.32%) (positive cla
ConfusionMatrix:
True: high low
high: 282 51
low: 59 608
f_measure: 91.71% +/- 1.32% (micro average: 91.70%) (positive
ConfusionMatrix:
True: high low
high: 282 51
low: 59 608</td><td>PerformanceVector

PerformanceVector:
accuracy: 94.10% +/- 2.28% (micro average: 94.10%)
ConfusionMatrix:
True: high low
high: 320 38
low: 21 621
classification_error: 5.90% +/- 2.28% (micro average: 5.90%)
ConfusionMatrix:
True: high low
high: 320 38
low: 21 621
kappa: 0.870 +/- 0.050 (micro average: 0.870)
ConfusionMatrix:
True: high low
high: 320 38
low: 21 621
lift: 146.88% +/- 3.75% (micro average: 146.78%) (positive clas
ConfusionMatrix:
True: high low
high: 320 38
low: 21 621
f_measure: 95.46% +/- 1.76% (micro average: 95.47%) (positive c
ConfusionMatrix:
True: high low
high: 320 38
low: 21 621</td></tr>
<tr><td>accuracy: 89.00% +/- 1.76% (micro average: 89.00%)
<table><tr><td></td><td>true high</td><td>true low</td></tr><tr><td>pred. high</td><td>282</td><td>51</td></tr><tr><td>pred. low</td><td>59</td><td>608</td></tr><tr><td>class recall</td><td>82.70%</td><td>92.26%</td></tr></table></td><td>accuracy: 94.10% +/- 2.28% (micro average: 94.10%)
<table><tr><td></td><td>true high</td><td>true low</td><td>clas</td></tr><tr><td>pred. high</td><td>320</td><td>38</td><td>89.3</td></tr><tr><td>pred. low</td><td>21</td><td>621</td><td>96.7</td></tr><tr><td>class recall</td><td>93.84%</td><td>94.23%</td><td></td></tr></table></td></tr>
</table>

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Va
Time_Employed	-0.294	-1.352	0.032	-9.2
Num_Prev_Positions	-1.749	-1.970	0.157	-11.
Teamwork	-0.781	-2.257	0.063	-12.
Motivation	-0.797	-2.306	0.065	-12.
Leadership	-0.744	-2.177	0.062	-11.
Intercept	19.400	2.087	1.359	14.2

B. Which model performed better and why? Which performance measures (list their values) were used to determine this and why?

Because the cost of both low_performace and high_perfomance are similar, Accuracy will be used to determine which model is better.

The decision tree model performed better because the key model performance metrics such as Accuracy, Kappa, Lift, and F-measure are higher for the decision tree model; therefore, they indicated that the decision tree analysis is performed better than the logistic regression.

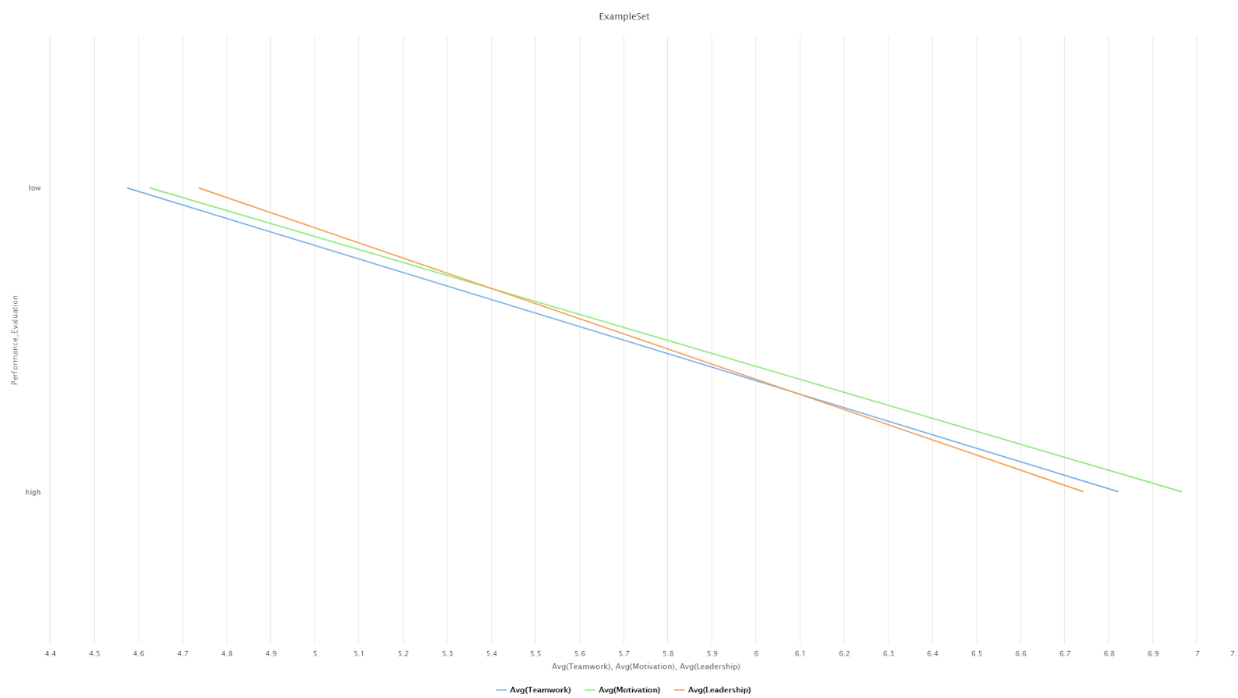
Metric	Decision Tree	Logistic Regression
Accuracy	94.10%	89% / 91.15%(pred. Low)
Kappa	.870	.754
Lift	146.88	138.47
F-measure	95.46	91.71

C. What business recommendations can be made after this analysis? Please write an Executive Summary (1 paragraph) including the following:

1. Note the business problem.
2. Briefly describe the steps taken to resolve the business problem.
3. Report the most important results (e.g., What factors are driving the outcome variable? What interesting insights does your model suggest?).
4. Briefly describe your business recommendations.

The firm would like to create a model to help predict the performance of potential new managers based on several variables: employment history and personality characteristics. The firm provided us with a data set of existing managers to build the model from. We have conducted data preparation of the dataset given by reviewing the data and cleaning it of inconsistencies, as well as removing personal identifiers that are unethical. After the data had been cleaned and prepared for analysis, we conducted three models: cluster analysis(unsupervised), decision tree, and logistic regression(supervised) to understand the various factors in what correlates to a manager's performance. In both supervised models, it is found that the three personality characteristics (motivation, teamwork, leadership) significantly determine the outcome of the performance of the manager. It is important to note that the longer a manager has been employed, the better their performance rating was. It is found that even with less experience in a management role that if the manager is rated high in the other factors, then there is a greater probability that they will perform high in their performance evaluation. It is our recommendation for the firm to first hire candidates that score high in these traits with experience, then hire those scoring high with less experience. It is also recommended that the firm employ leadership training for current employees and increase employee engagement to help lift the scores of leadership, teamwork, and motivation.

D. Create at least one visualization to demonstrate the most salient point(s) from your analysis and/or recommendations and provide it here. This visualization should support something reported in 5C. Include a caption.



This chart shows the aggregated average data of Teamwork, Motivation, and leadership in relation to performance evaluations. As seen in the graph, the higher the scores in the personality metrics, the higher the probability that the performance evaluation was in the high category.