Data Mining Final Project Report by Abhishek Wagh and Pradeep Thiagu

Introduction:

In the final project, we want to apply the concepts we learnt in class to real life scenarios. We assumed that we are consultants working in a company and the company's director has approached us for a task. The task is to give the director recommendations to help improve the working of certain departments. The director will then forward the recommendations to the concerned departments. The director has given us 4 datasets regarding human resources (HR) issues, service agent, sales, and customer churn. We must analyze, mine and then present our recommendations to the director.

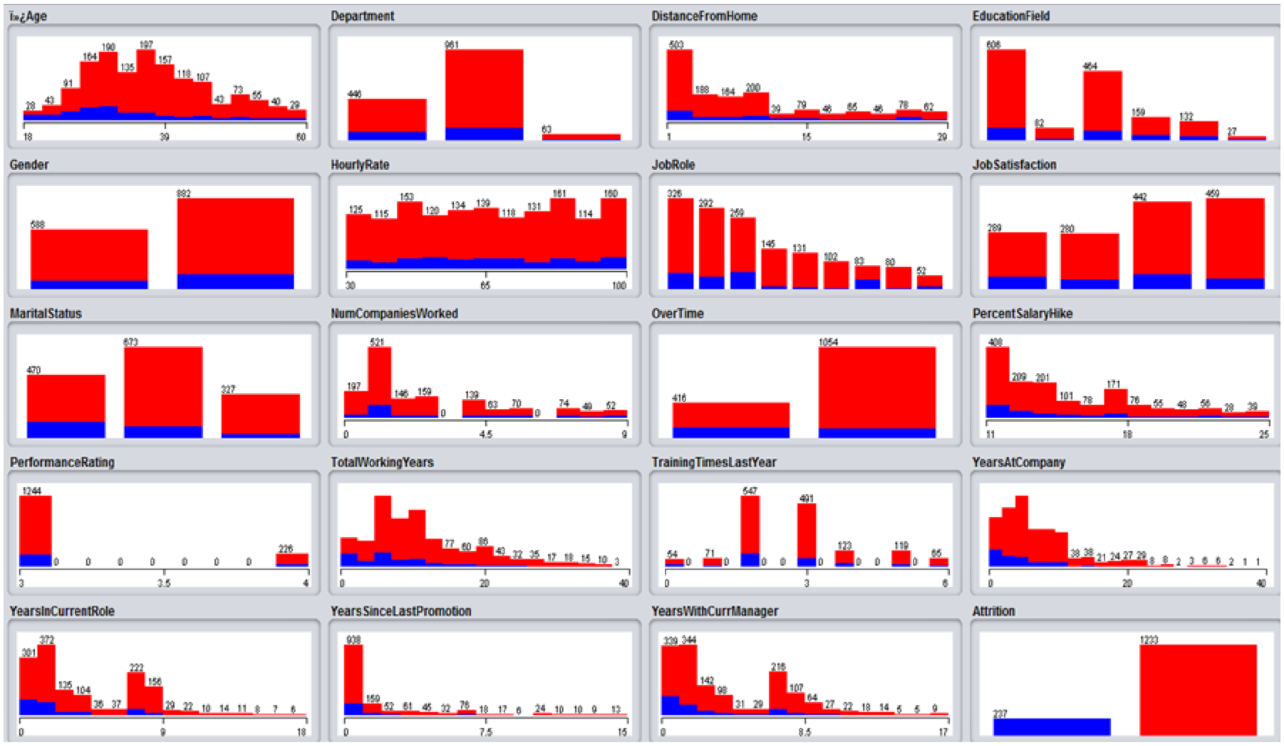
Methodology: Generally, we first used Weka to do the initial analysis of the datasets. We first find out if the methods we learned in class can be used on other datasets. We will do analysis on Weka such as classification, clustering, boundary visualization. We will compare the accuracy of Weka with the accuracy of dataset in IBM Watson. We then analyze the dataset in IBM Watson to get a graphical representation of the datasets which can then be presented to the director. We will then recommend certain steps which the company can take to improve performance of the departments. For some datasets, we used different approaches based on whether we can get better results and as a part of learning through experimentation.

Dataset 1: Human Resources

We uploaded the HR-employee-attrition dataset on the Weka preprocess page. We found that the dataset has 1471 instances and 35 attributes. The number of attributes is quite large. But according to our experience after working in a company for five years, we notice that some attributes are more significant compared to other attributes and so we removed some attributes and we selected 20 attributes from the 35 attributes. The following are the attributes which we selected:

Age, attrition, department, distance from home, educational field, gender, hourly rate, job role, job satisfaction, marital status, number of companies worked, overtime, percentage salary hike, performance rating, total working years, training times last year, years in company, years in current role, years since last promotion, years with current manager.

Of the 20 attributes, we wanted to find with what accuracy we can predict whether an employee will leave the company or not (attrition). Since attrition is one of the attributes we selected, we wanted to find whether the other 19 attributes can correctly predict whether attrition will occur or not. So, we went to the edit tab in the preprocess page and we selected attrition to be the class. We then get the following plots:



We then went to the classification page to perform two classifications which we think will give a more accurate classification: NaiveBayes and Lazy IBk classification. We initially use NaiveBayes as a classifier to check the accuracy of the classifier. We select 10-fold cross-validation and we find that the correctly classified instances is 82.8571% which is quite good. It means that the 19 attributes we selected can correctly predict attrition 82.8571% of the time.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1218               82.8571 %
Incorrectly Classified Instances       252               17.1429 %
Kappa statistic                          0.3871
Mean absolute error                      0.2377
Root mean squared error                  0.3579
Relative absolute error                 87.7849 %
Root relative squared error             97.3336 %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.511    0.110    0.471      0.511    0.490      0.388    0.757     0.488     Yes
                 0.890    0.489    0.904      0.890    0.897      0.388    0.757     0.916     No
Weighted Avg.    0.829    0.428    0.834      0.829    0.831      0.388    0.757     0.847

=== Confusion Matrix ===

    a    b   <-- classified as
  121  116 |    a = Yes
  136 1097 |    b = No
```

We select another classifier which is the Lazy IBk classifier to test the accuracy of the dataset. We find that the 19 attributes we selected can correctly predict whether a person will leave the company 78.7755% of the time. The accuracy percentage is quite high in both the lazy IBK and the NaiveBayes methods.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1158               78.7755 %
Incorrectly Classified Instances       312               21.2245 %
Kappa statistic                          0.1485
Mean absolute error                      0.2127
Root mean squared error                  0.4604
Relative absolute error                 78.5342 %
Root relative squared error            125.184  %
Total Number of Instances             1470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.245    0.108    0.304      0.245    0.271      0.150    0.568     0.196     Yes
                 0.892    0.755    0.860      0.892    0.876      0.150    0.568     0.858     No
Weighted Avg.    0.788    0.651    0.770      0.788    0.778      0.150    0.568     0.751

=== Confusion Matrix ===

    a    b   <-- classified as
   58  179 |    a = Yes
  133 1100 |    b = No
```

We then went to the clustering page to perform clustering on the dataset and find out whether clustering (unsupervised classification) will give similar results to supervised classification:
Cluster analysis:

Cluster using EM:

```
Clustered Instances

0         636 ( 43%)
1         834 ( 57%)
```

Cluster using Simple K means:

```
Clustered Instances

0         827 ( 56%)
1         643 ( 44%)
```

Using cluster analysis, we take the number of clusters as 2 and we do cluster analysis using 2 clustering methods: EM and simple K means and we get the results as described above. We find that the result of clustering is quite different from the result of classification since the clustering classification is unsupervised. We also find that by using different clustering methods, we get different results for whether attrition has taken place or not. The result of clustering using EM is - attrition is Yes: 834, attrition is no: 636. The result of clustering using Simple K-Means is - attrition is Yes: 643, attrition is no: 827. But in reality (in dataset), attrition is Yes:237, attrition is No: 1233. So, the accuracy of cluster analysis is quite inaccurate due to large difference between attrition in reality and attrition using cluster analysis. However Simple K-Means has result closer to the real attrition, so Simple K-Means method can be more accurate in analyzing this dataset using clustering than EM method.

We wanted to find a formula with which we could predict the number of years an employee may remain in the company based on the person's age and other performance parameters. We removed all the nominal attributes and we only considered the numerical attributes such as age, number of companies in which the employee has worked, percentage salary hike, total number of working years, years in current role, years since last promotion and years with current manager. We selected the attribute years in company as class. We went to the classification page and did linear regression on the dataset which only has the numerical attribute which we selected. After analyzing the dataset using linear regression, we got the following equation:

(-0.0323 * Age) + (-0.2923 * NumCompaniesWorked)+

(-0.0367 * PercentSalaryHike) + (0.2716 * TotalWorkingYears) +

(0.4657 * YearsInCurrentRole) + (0.3187 * YearsSinceLastPromotion) +

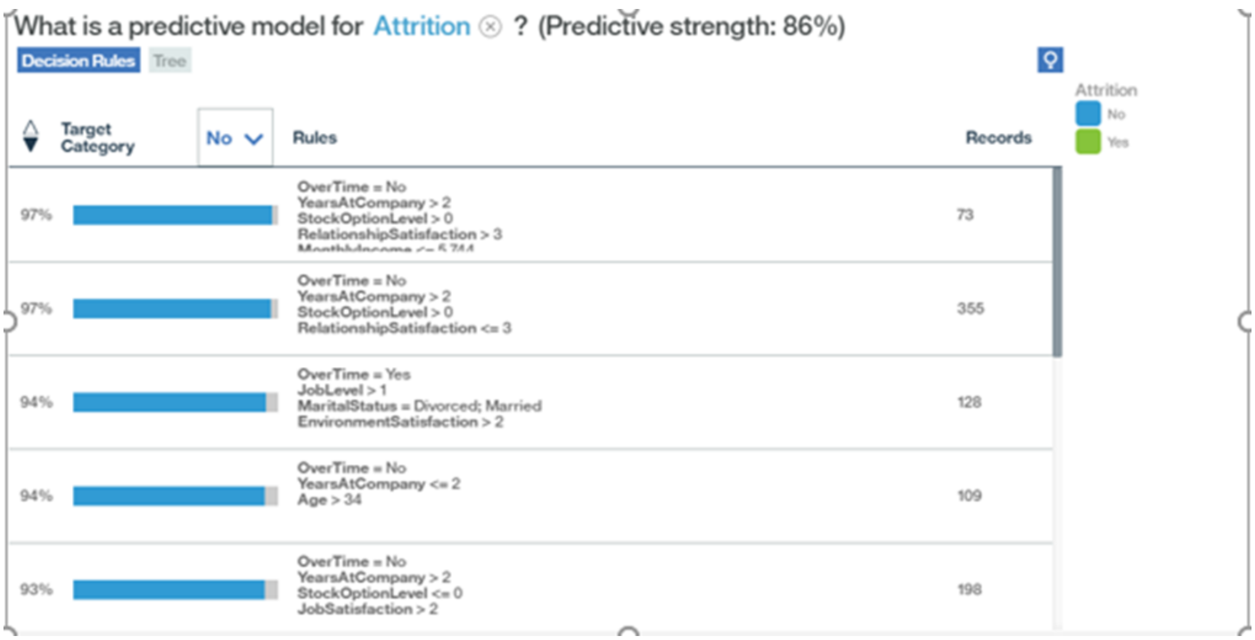
(0.5578 * YearsWithCurrManager) + 1.5157

Thus, by using the above equation and substituting the numerical values of the attributes we will get the estimated value of the number of years an employee may work in the company.

We want to present a graphical analysis to the director. We came to know that IBM Watson can provide visualization of the various aspects of the attributes. So, we setup an account on IBM Watson and did the dataset analysis on the website. IBM Watson applies a multitude of Machine Learning Algorithms to get the best analytics results. We have displayed the results of the analysis below:

Predictive model of attrition: Yes

The above model shows that the main factors for attrition is because of overtime, job level is <=1 which means the job is quite easy and the employee does not have stock options. The HR department needs to address these three factors to retain employees.



The above predictive model shows that the best way the HR can best predict (with 97% accuracy) whether an employee may not leave the company is when the employee does not work overtime, the employee is in the company for more than 2 years, the employee has at least 1 stock options, has good relationship satisfaction which is greater than 3.

The highest accuracy we got in Weka was 82.8571% by using the NaiveBayes method. The accuracy using IBM Watson is 86%. Both the percentage accuracies are quite close and we can conclude that the model created by the dataset is quite accurate.

What is a predictive model for PercentSalaryHike ? (Predictive strength: 60%)

Decision Rules Tree

Predicted value	Rules	Records
22.23	PerformanceRating = 4 YearsAtCompany > 7	88
21.61	PerformanceRating = 4 YearsAtCompany <= 7	138
14.43	PerformanceRating = 3 YearsAtCompany <= 2	295
13.87	PerformanceRating = 3 YearsAtCompany > 2	949

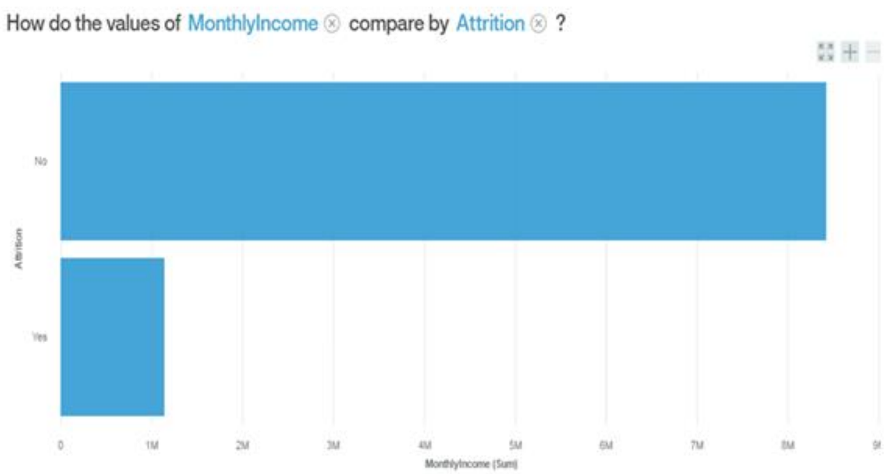
The above model shows the predictive model for percent salary hike. The predictive strength is 60%. The model shows that the best predictor that a person will get a salary hike is if the performance rating = 4 and the person is in company for more than 7 years.

What is a predictive model for YearsAtCompany ? (Predictive strength: 79%)

Decision Rules Tree

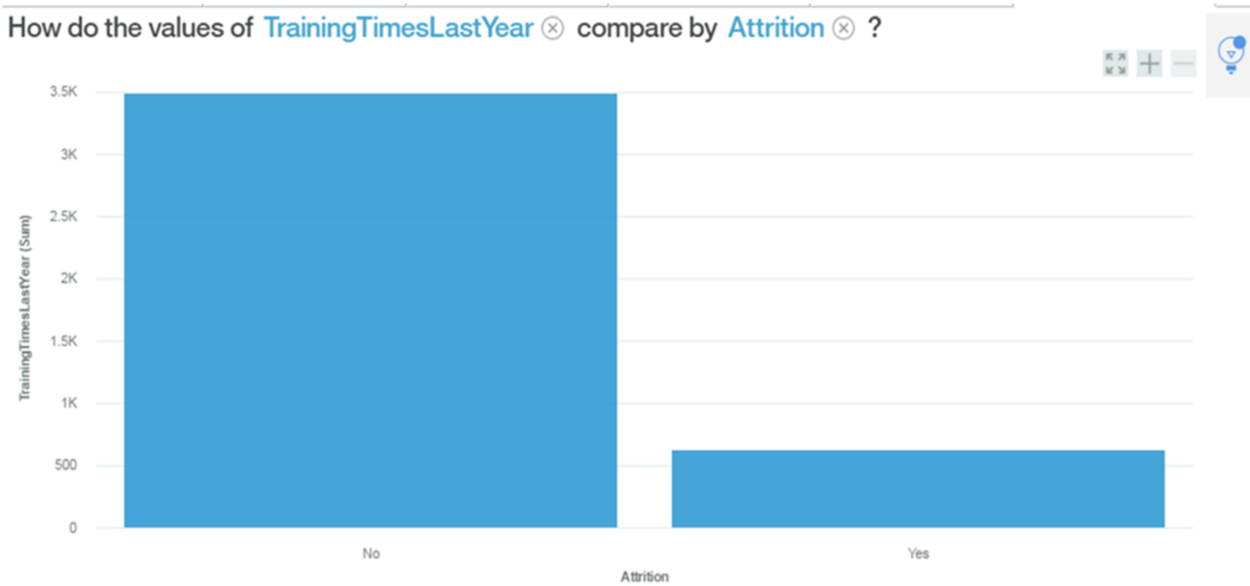
Predicted value	Rules	Records
23.64	YearsWithCurrManager > 7 TotalWorkingYears > 17 NumCompaniesWorked <= 1	50
17.83	YearsWithCurrManager > 7 TotalWorkingYears > 17 NumCompaniesWorked > 1	65
15.58	YearsWithCurrManager = 4 to 7 TotalWorkingYears > 17	52
12.78	YearsWithCurrManager > 7 TotalWorkingYears = 10 to 17	87
9.65	YearsWithCurrManager > 7 TotalWorkingYears = 8 to 10	69

The above predictive model shows what factors may help the HR determine whether an employee will stay with the company for a longer time. The predictive model shows that the best factors are that the employee stays with current manager for more than 7 years, the employee has worked totally for more than 17 years and that the employee has worked for only 1 company till now i.e. the employee has not changed companies. This model shows that an employee who stays with a manager for a longer time, has worked totally for a long time (thus age > 35 years (18 + 17 years) and has not changed companies will stay in the company for a longer time.

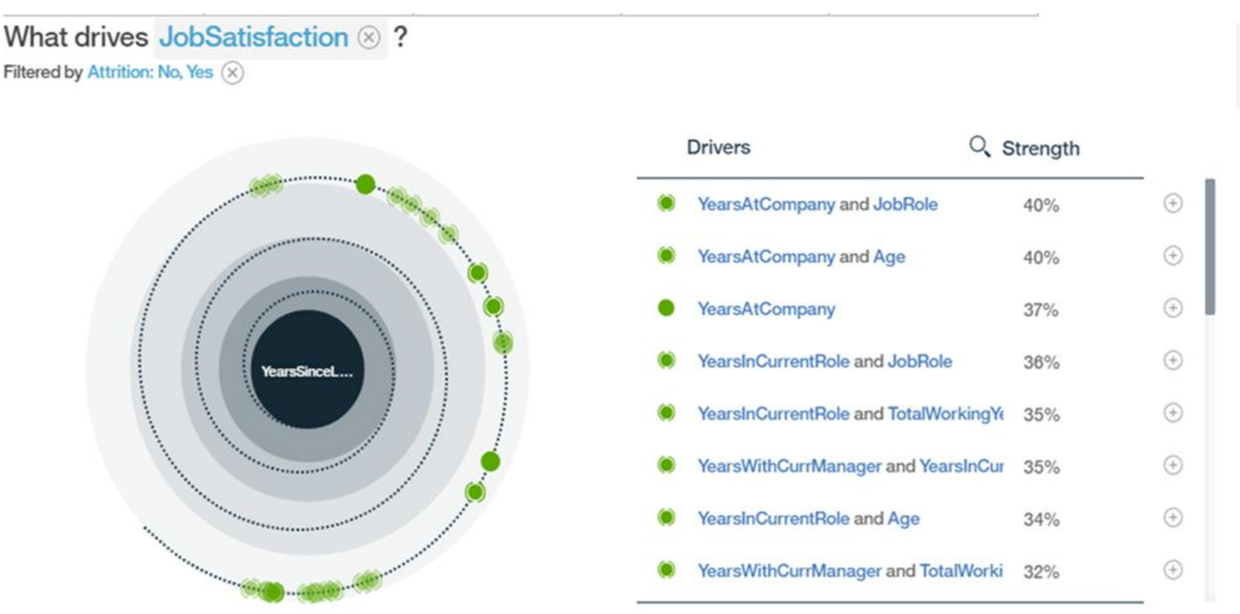


The above graph shows that how attrition varies with monthly income. It shows that as the monthly income increases, the attrition reduces.

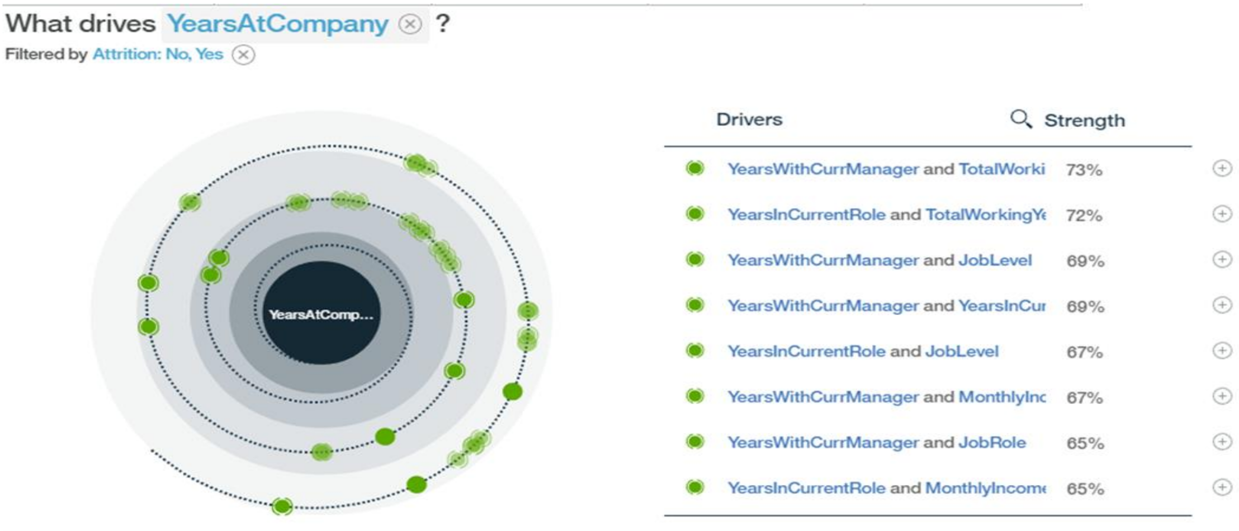
The above graph shows the relationship between percentage salary hike versus attrition. The plot shows that as attrition rate reduces, the salary hike increases. Thus, it is essential for the HR department to recommend decent salary hikes to employees to reduce attrition.



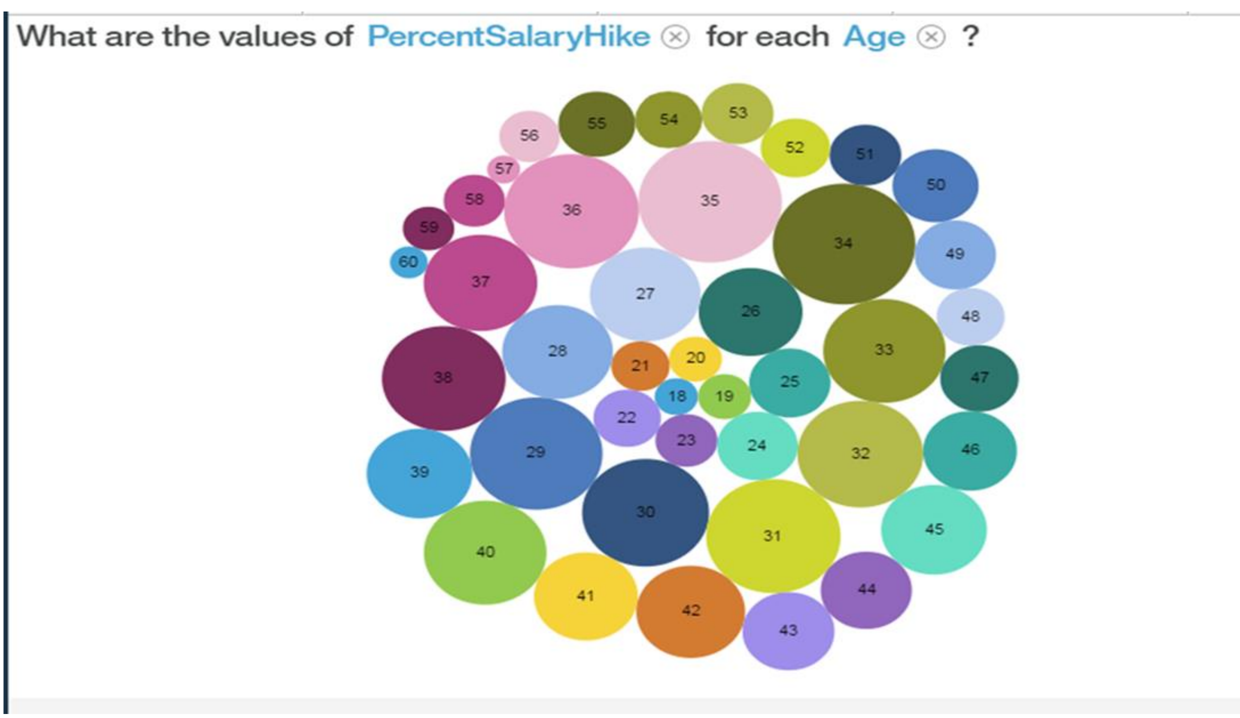
This graph shows the relationship between attrition and training time last year. The graph means that if the number of times an employee gets trained increases, the attrition reduces. Thus, HR must ensure that the employee should get trained at more regular intervals or else the employee's job skills may become redundant.



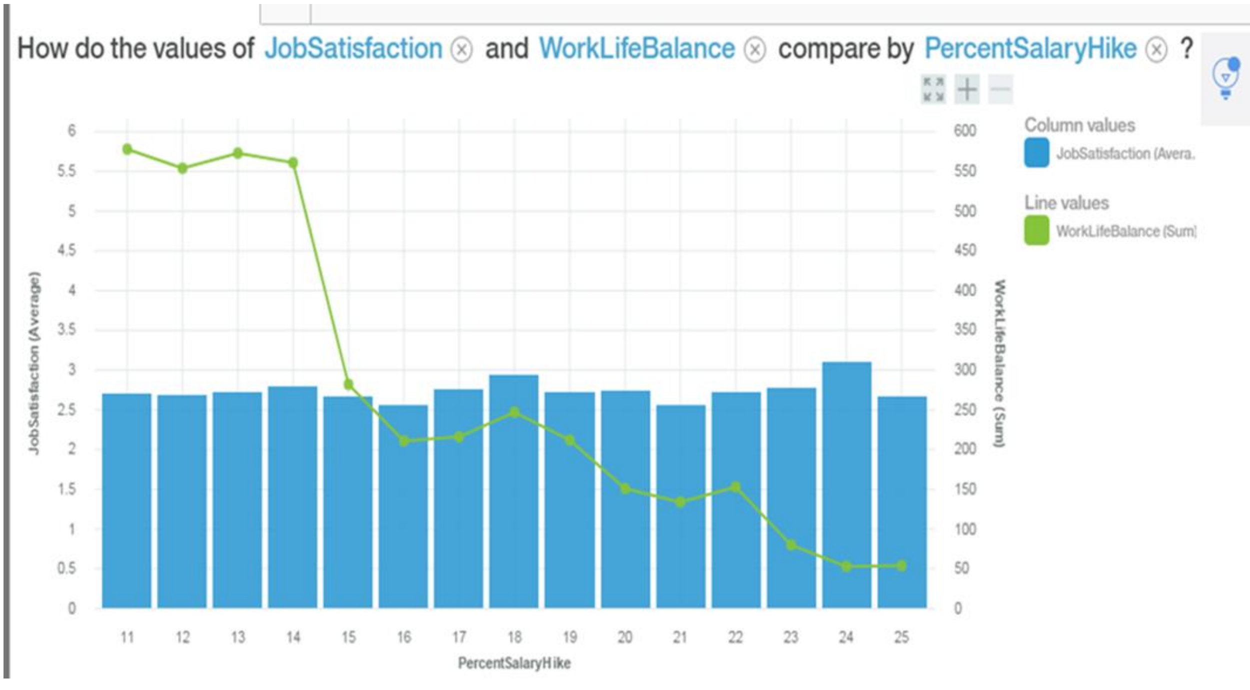
The above figure shows the drivers of job satisfaction in the company. The most important drivers are Job Role, years at company and age, followed by years in current role, total working years and years with current manager. All these mentioned factors must be considered by the HR department to improve an employee's job satisfaction.



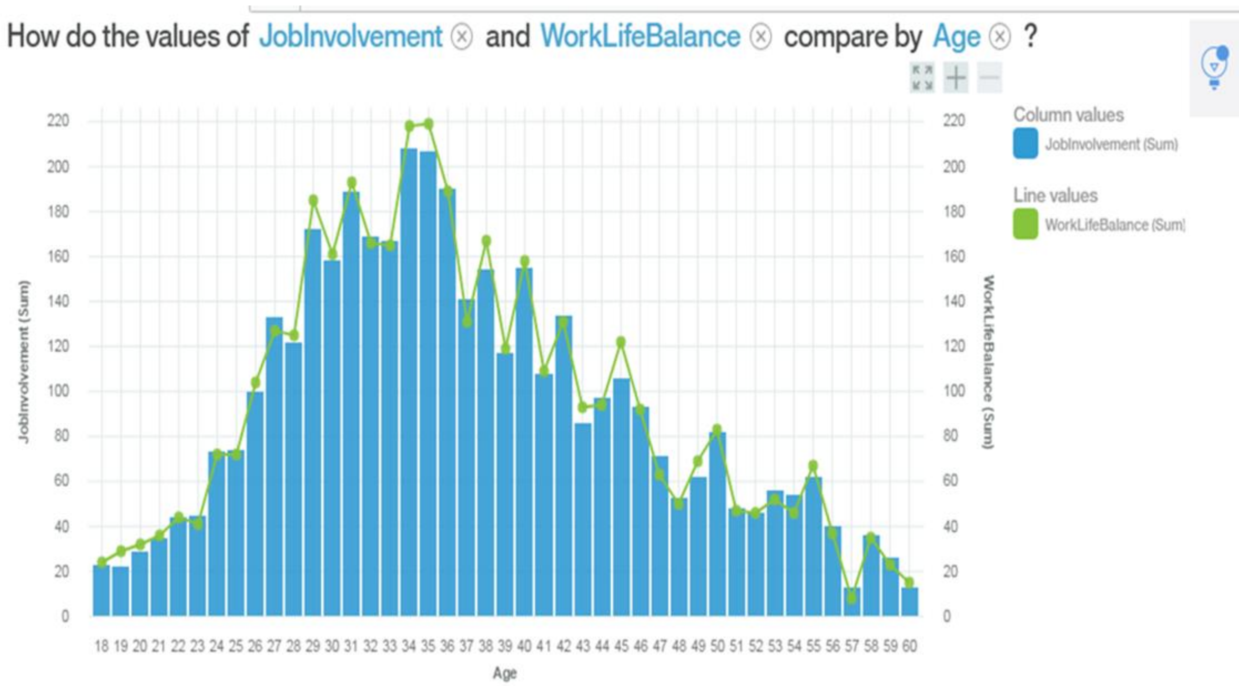
The above figure shows the factors which drive the number of years an employee is with the company. The most important factors are years with current manager, total working years, years in current role, job level and monthly income. Thus, if the HR department sees that many employees are leaving a department, it is possible that the managers may not be treating the employees well. Thus, if the company wants to retain employees for a certain amount of years to recover the investment made in an employee, the HR department should look into the main factors that drive the number of years an employee is with the company.



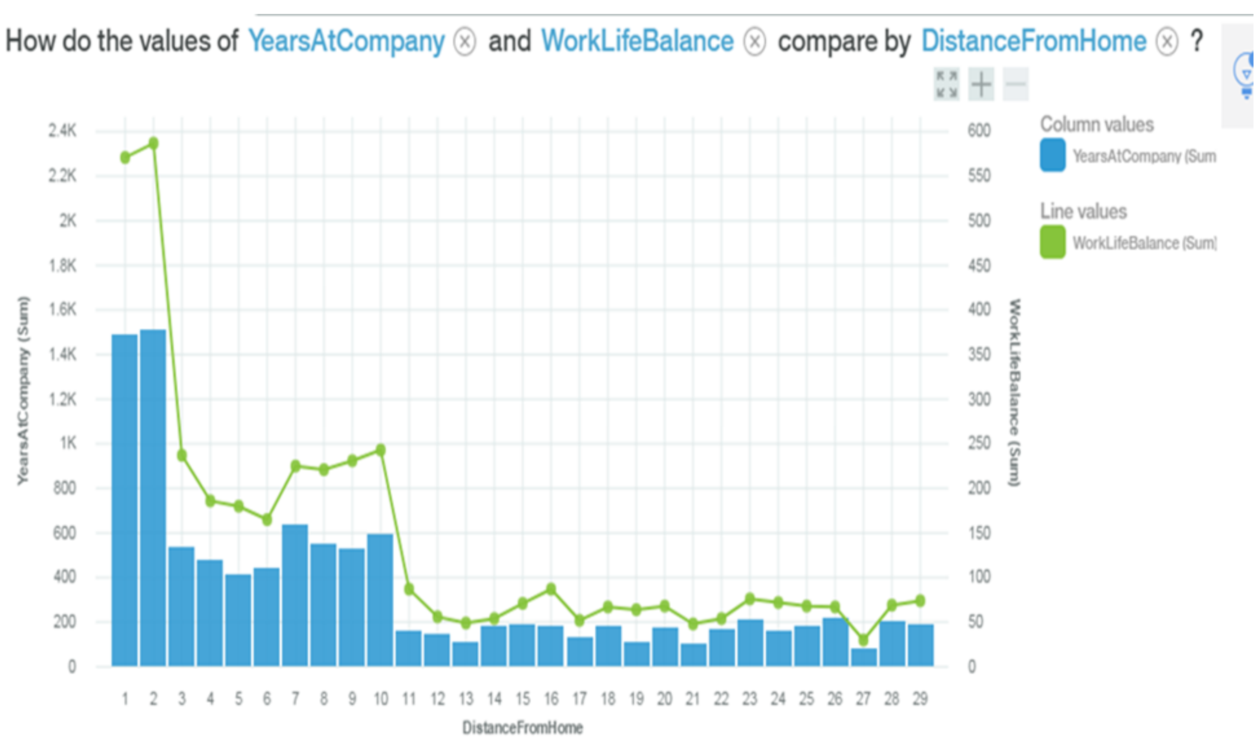
This figure shows the percentage salary hike for each age. The number in each circle resembles the age and the size of the circles represents the size of the salary hike. A larger circle size means a larger salary hike. The figure shows that most of the salary hikes are significant from the ages of 26 to 42. Thus, people in that age group are more productive as they are experienced in their jobs and have the energy to perform. This looks like a disturbing trend since the older employees who have worked for many years are not given significant salary hikes. Thus, the HR should formulate policies which can give significant salary hikes to employees who are older than 42, are more experienced and have high performance ratings.



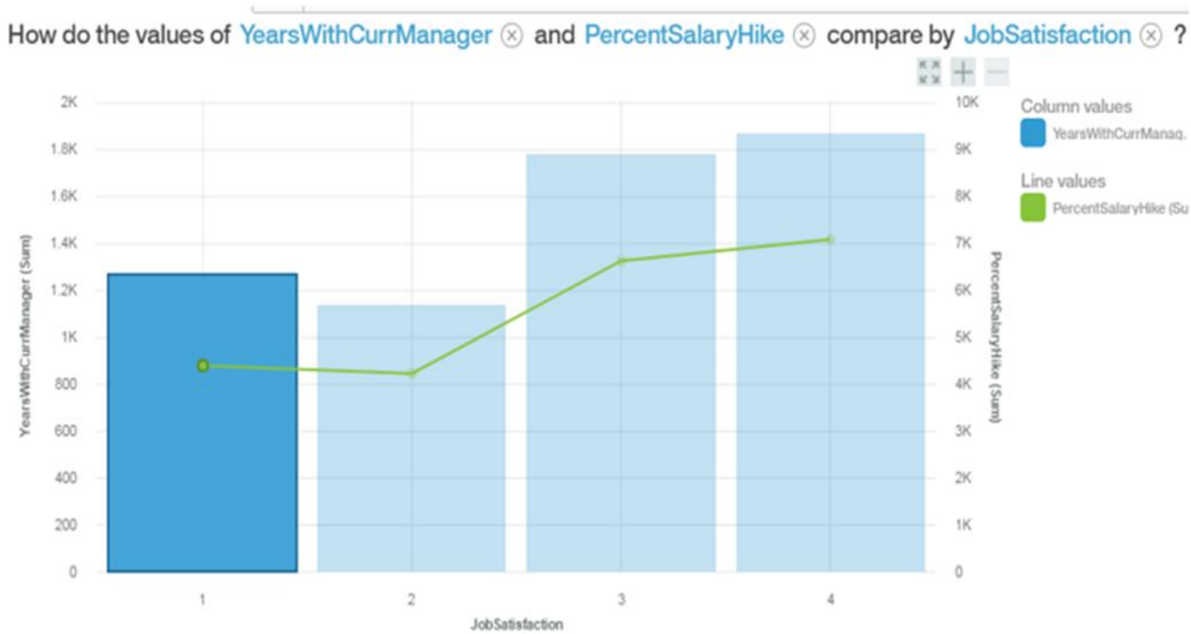
The above figure shows that how percentage salary hike affects job satisfaction and work life balance. We can see that if salary is hiked, there is minimal noticeable change in job satisfaction but there is a drastic change in the work life balance as the salary hike becomes more than 14%. We can conclude that a salary hike will not increase a person's job satisfaction, but work-life balance will change because a person will work much harder, will take more work load or may take tougher jobs which may take more time. Thus, more time taken in work may increase stress in employees, thus affecting work life balance.



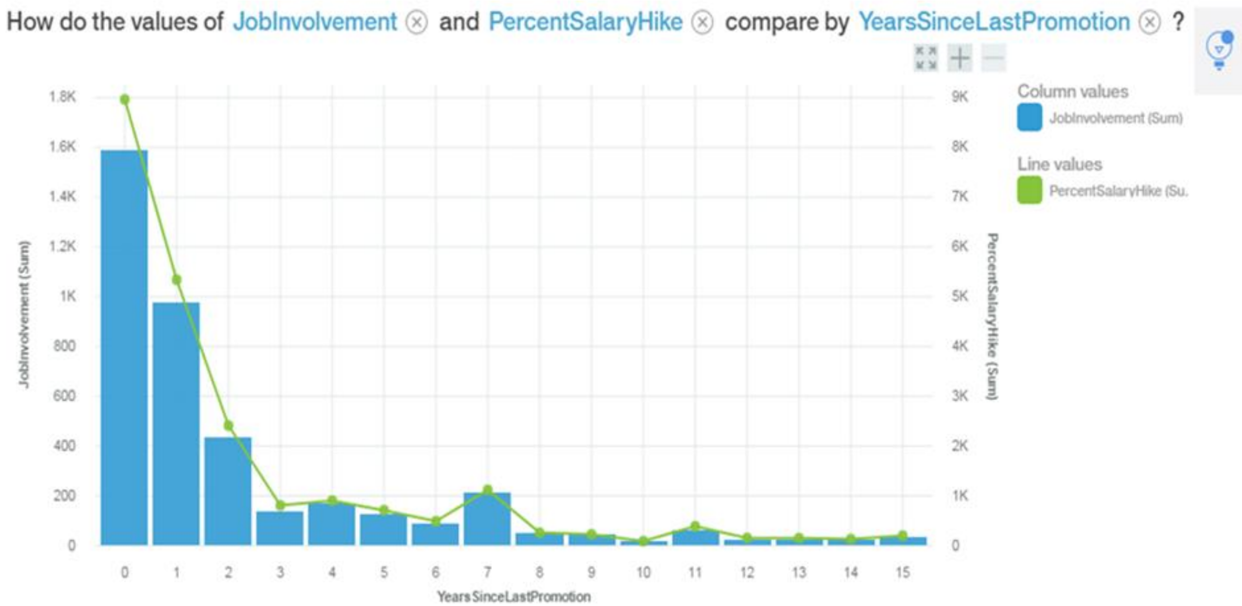
This graph shows that how age affects the job involvement and work -life balance. The analysis shows that a person is most satisfied with job and work-life balance between the ages of 27 to 42 years. Also, if a person is satisfied with a job, then work-life balance also improves. Thus, the HR should try to improve job involvement of employees from outside the range of age of 27 to 42 years to keep employees more content by examining the reasons for poor job involvement and addressing the issues proactively.



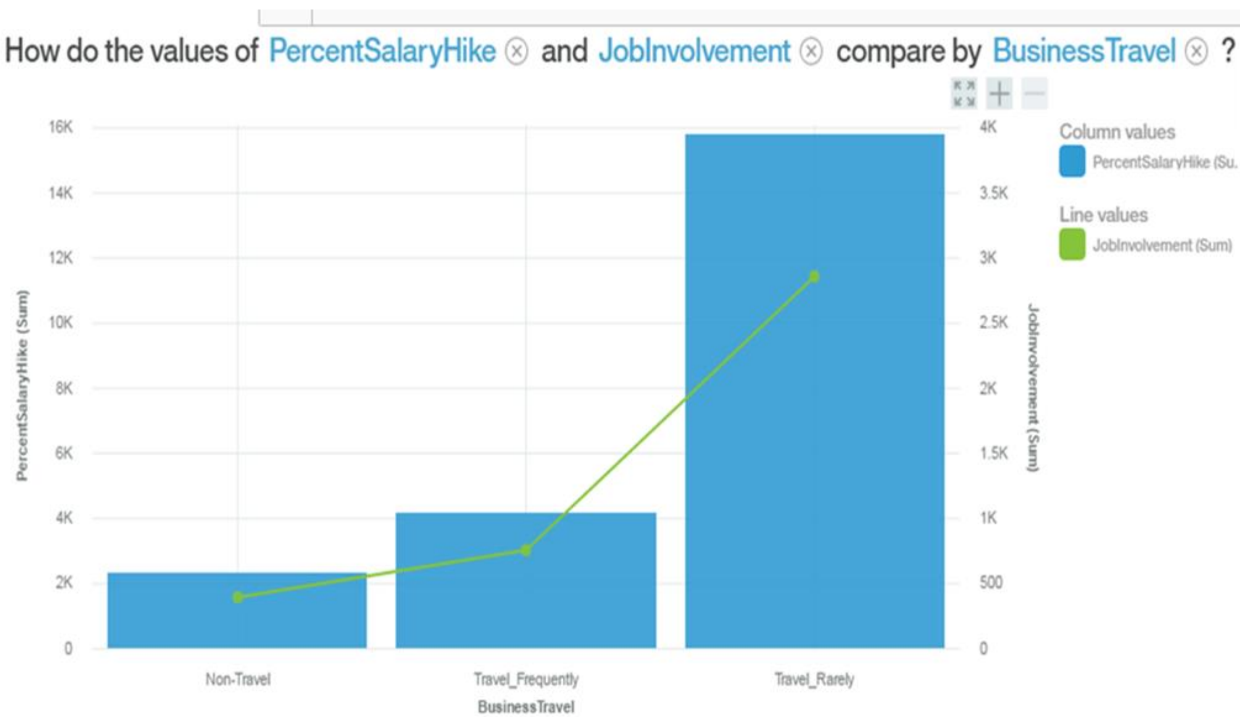
This graph shows the effect of distance from home on work-life balance and years at company. It shows that as a person stays further away from company, the work life balance and years at the company reduces. This shows that as distance from home increases, a person is less content with the job. So, work-life balance reduces and the person tries to find a new job faster. Thus, the HR department should ask management to increase housing allowance to employees or build a residential complex near the office so that the employees can stay nearer to the office.



This graph shows the variation of percentage salary hike and years with current manager with respect to job satisfaction. It shows that as job satisfaction increases, the employee will stay with current manager for a longer time. Also, since the employee performs better in job, the employee will get better performance ratings and thus the employee's percentage salary hike will also increase. Thus, the HR department should take steps to improve employee's job satisfaction such as holding training sessions for employees regarding how to deal with other employees so as to improve job satisfaction.



This graph shows the effect the years since last promotion has on job involvement and percentage salary hike. The graph shows that if years since the last promotion increases, job involvement reduces and percentage salary hike reduces. This effect is because once an employee gets promoted, the employee gets a large salary hike due to change in position. Due to salary hike, the employee is more involved in the job due to extra zealousness. As the time goes by, the employee gets less involved in the job and thus the employee gets less percentage salary hike due to poorer performance reviews. The HR (Human Resources) department should try to formulate a policy in which employees can improve job involvement after they are promoted.



This plot shows that how percentage salary hike and job involvement is related to business travel. The plot shows if an employee travels rarely, the salary hike is much more than travelling frequently or not travelling at all. Also, job involvement is most when employee is travelling rarely. Thus, if the company wants to improve job involvement of its employees, it should encourage employees to travel a few business related trips a year. Travelling will ensure that employees get some exposure to new ideas and travelling less ensures that more employees get a chance to travel, thus job involvement of more employees will improve. Travelling frequently causes employees to get tired and so reduces job involvement.

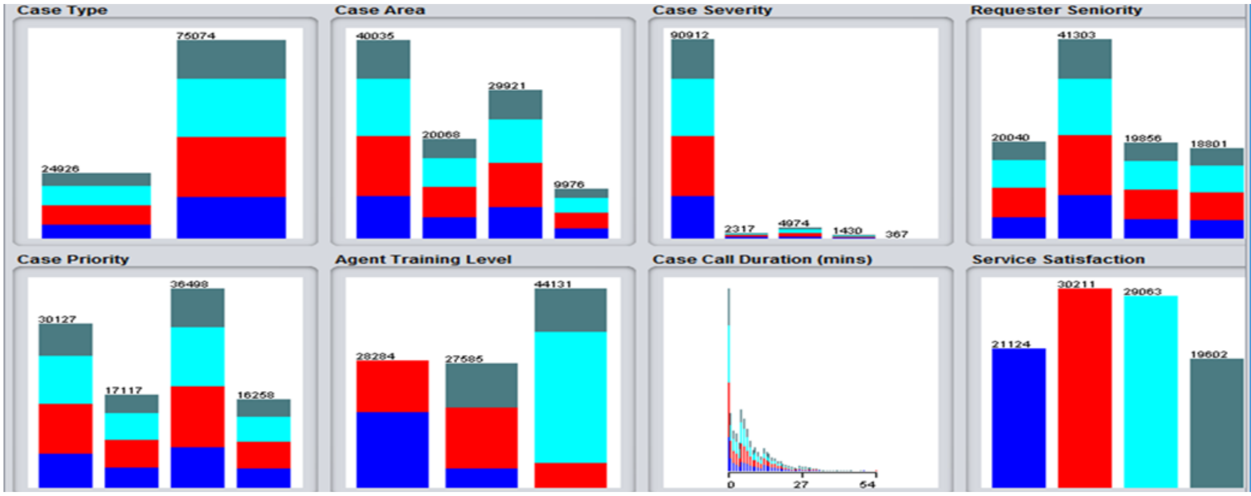
Dataset 2:

Service agent performance dataset:

We uploaded the service agent dataset to Weka, we selected service satisfaction attribute as class. The dataset has 100,000 instances and 8 attributes.

Case #	Case Type	Case Area	Case Severity	Requester ID	Requester Seniority	Case Priority	Service Agent ID	Agent Training Level	Case Call Duration (mins)	Service Satisfaction
1	Issue	Systems	2_Normal	1929	1_Junior	0_Unassigned	50	0_No training	3	1_Unsatisfied
2	Request	Software	1_Minor	1587	2_Regular	1_Low	15	0_No training	5	1_Unsatisfied
3	Request	Access/Login	2_Normal	925	2_Regular	0_Unassigned	15	1_Minimal training	0	0_Unknown
4	Request	Systems	2_Normal	413	4_Management	0_Unassigned	22	1_Minimal training	20	0_Unknown
5	Request	Access/Login	2_Normal	318	1_Junior	1_Low	22	0_No training	1	1_Unsatisfied
6	Request	Access/Login	2_Normal	858	4_Management	3_High	38	1_Minimal training	0	0_Unknown
7	Request	Systems	2_Normal	1978	3_Senior	3_High	10	1_Minimal training	9	0_Unknown
8	Request	Software	2_Normal	1209	4_Management	0_Unassigned	1	1_Minimal training	15	0_Unknown
9	Request	Software	2_Normal	887	2_Regular	2_Medium	14	0_No training	6	1_Unsatisfied
10	Request	Access/Login	2_Normal	1780	3_Senior	1_Low	46	0_No training	1	1_Unsatisfied
11	Request	Software	2_Normal	1349	3_Senior	3_High	1	1_Minimal training	7	0_Unknown
12	Request	Systems	2_Normal	1893	2_Regular	1_Low	50	0_No training	17	1_Unsatisfied
13	Request	Systems	2_Normal	645	2_Regular	3_High	11	0_No training	10	1_Unsatisfied
14	Issue	Systems	2_Normal	1492	4_Management	3_High	26	2_Sufficient training	4	3_Highly satisfied
15	Issue	Software	2_Normal	1978	3_Senior	0_Unassigned	9	2_Sufficient training	7	3_Highly satisfied
16	Request	Software	2_Normal	216	4_Management	0_Unassigned	7	2_Sufficient training	11	2_Satisfied
17	Request	Access/Login	2_Normal	1586	2_Regular	0_Unassigned	20	2_Sufficient training	0	3_Highly satisfied
18	Request	Systems	2_Normal	1554	2_Regular	2_Medium	42	2_Sufficient training	7	3_Highly satisfied
19	Request	Systems	2_Normal	518	4_Management	0_Unassigned	16	2_Sufficient training	7	3_Highly satisfied

We remove attributes case call number, requester ID and service agent ID for de identification purposes. We get the following plots as shown below:



These plots show the quantity of categories of nominal attributes case types, case area, case severity, requester seniority, case priority, agent training level and service satisfaction. It also shows maximum, minimum, mean and standard deviation of numerical attribute which is case call duration.

We then go to the classification page and to do classification, we use Naive Bayes classification and we find that the accuracy of dataset is 58.697%.

```
Correctly Classified Instances       58697               58.697  %
Incorrectly Classified Instances     41303               41.303  %
Kappa statistic                          0.4338
Mean absolute error                      0.2622
Root mean squared error                  0.3644
Relative absolute error                 70.7469 %
Root relative squared error             84.651  %
Total Number of Instances           100000

     a     b     c     d   <-- classified as
 16406  4717     0     1 |     a = 1_Unsatisfied
 11126 13617  5251   217 |     b = 0_Unknown
    85     0 28274   704 |     c = 3_Highly satisfied
   254  9662  9286   400 |     d = 2_Satisfied
```

We then use J48 classification to see if we can get a better classification accuracy and we were successful. We got an accuracy of 59.26% using the J48 classification.

J48 Classification:

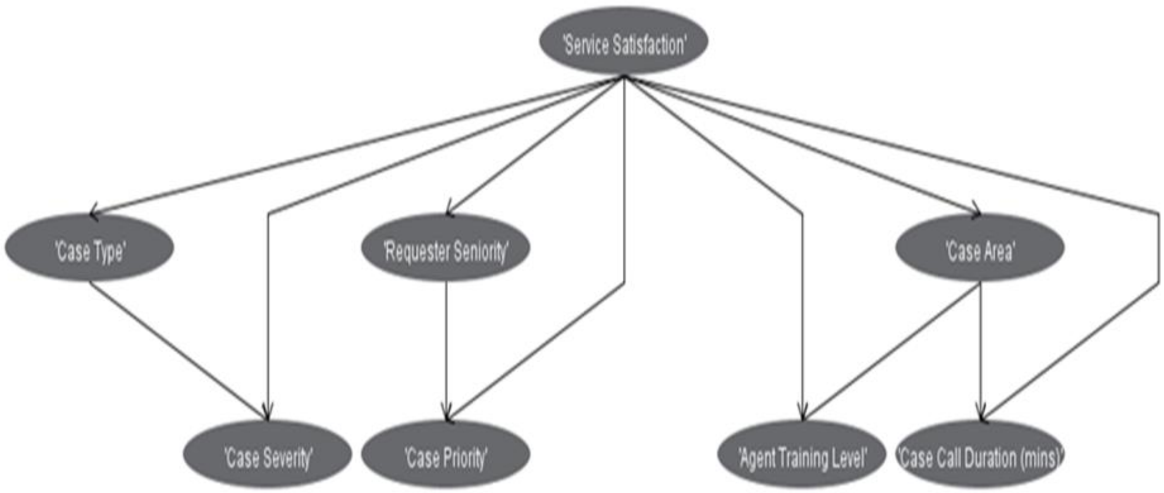
```
Correctly Classified Instances       59260               59.26   %
Incorrectly Classified Instances     40740               40.74   %
Kappa statistic                          0.4375
Mean absolute error                      0.2629
Root mean squared error                  0.3632
Relative absolute error                 70.9319 %
Root relative squared error             84.3721 %
Total Number of Instances           100000
```

```
     a      b      c      d   <-- classified as
 15827   5293      0      4 |     a = 1_Unsatisfied
 10282  14444   5467     18 |     b = 0_Unknown
    85      0  28978      0 |     c = 3_Highly satisfied
    87   9826   9686      3 |     d = 2_Satisfied
```

We wanted to explore the relation between the attributes. So, we selected BayesNet and chose the maximum number of parents as 2. We did BayesNet analysis using maximum 2 parents per attribute and we got accuracy as 59.252%. We did not select maximum 3 or more parents per attribute since the results of the probability distribution may become more complex and will be difficult to display to the director.

```
Correctly Classified Instances       59252               59.252 %
Incorrectly Classified Instances     40748               40.748 %
Kappa statistic                          0.4394
Mean absolute error                      0.2619
Root mean squared error                  0.3621
Relative absolute error                 70.6703 %
Root relative squared error             84.1082 %
Total Number of Instances           100000
     a     b     c     d   <-- classified as
 15827  5293     0     4 |     a = 1_Unsatisfied
 10282 14444  5467    18 |     b = 0_Unknown
    85     0 28978     0 |     c = 3_Highly satisfied
    87  9826  9686     3 |     d = 2_Satisfied
```

After using BayesNet with maximum number of parents as 2, we then get the visualization of the attribute and their parents.



We then see the probability distribution table of Agent training level since we want to see how much training level is required so that the agents are competent to troubleshoot problems.

Probability Distribution Table For 'Agent Training Level'

'Service Satisfaction'	'Case Area'	'0_No training'	'1_Minimal training'	'2_Sufficient training'
1_Unsatisfied	Systems	0.788	0.212	0
1_Unsatisfied	Software	0.797	0.203	0
1_Unsatisfied	Access/Login	0.795	0.205	0
1_Unsatisfied	Hardware	0.805	0.195	0
0_Unknown	Systems	0.373	0.447	0.18
0_Unknown	Software	0.371	0.444	0.185
0_Unknown	Access/Login	0.372	0.446	0.182
0_Unknown	Hardware	0.404	0.421	0.175
'3_Highly satisfied'	Systems	0	0	1
'3_Highly satisfied'	Software	0	0	1
'3_Highly satisfied'	Access/Login	0	0	1
'3_Highly satisfied'	Hardware	0.029	0	0.971
2_Satisfied	Systems	0	0.5	0.5
2_Satisfied	Software	0	0.516	0.484
2_Satisfied	Access/Login	0	0.504	0.495
2_Satisfied	Hardware	0.045	0.467	0.489

We can interpret the table by this example:
The probability that the agent has no training given that service satisfaction is unsatisfied and that the case area is hardware is 0.805.
We see from the above table that service satisfaction was “highly satisfied” in all four case areas if the agents were sufficiently trained and service satisfaction was poor if agents were not trained. Thus, the company should invest in sufficient training of the service agent will the agents.

Call duration probability distribution table:

'Service Satisfaction'	'Case Area'	'\'(-inf-0.5]\''	'\'(0.5-2.5]\''	'\'(2.5-7.5]\''	'\'(7.5-12.5]\''	'\'(12.5-inf)\''
1_Unsatisfied	Systems	0	0.06	0.328	0.141	0.471
1_Unsatisfied	Software	0.014	0.121	0.437	0.277	0.151
1_Unsatisfied	Access/Login	0.707	0.28	0.012	0	0
1_Unsatisfied	Hardware	0	0	0.062	0.236	0.702
0_Unknown	Systems	0	0.084	0.439	0.161	0.316
0_Unknown	Software	0.023	0.161	0.524	0.2	0.092
0_Unknown	Access/Login	0.785	0.208	0.007	0	0
0_Unknown	Hardware	0	0	0.095	0.356	0.548
'3_Highly satisfied'	Systems	0	0.108	0.514	0.181	0.197
'3_Highly satisfied'	Software	0.029	0.191	0.589	0.131	0.062
'3_Highly satisfied'	Access/Login	0.844	0.152	0.005	0	0
'3_Highly satisfied'	Hardware	0	0	0.137	0.443	0.419
2_Satisfied	Systems	0	0.085	0.417	0.177	0.321
2_Satisfied	Software	0.02	0.154	0.523	0.2	0.102
2_Satisfied	Access/Login	0.792	0.202	0.005	0	0
2_Satisfied	Hardware	0	0	0.098	0.333	0.569

From this table, we conclude that in systems and software cases, the issues are moderately difficult to solve since on average, they take between 2.5 minutes to 7 minutes to solve. In the case of login, most cases can be solved in 30 seconds, which means that the login cases are quite easy to solve. But in the case of hardware, the call durations are quite long in case of calls classified as satisfied and highly satisfied. This shows that hardware cases are hard to solve. So, training to the service agents needs to be enhanced in the case of hardware due to the difficulty in solving the hardware cases.

Probability Distribution Table For 'Case Priority'

'Service Satisfaction'	'Requester Seniority'	0_Unassigned	1_Low	3_High	2_Medium
1_Unsatisfied	1_Junior	0.292	0.402	0.104	0.202
1_Unsatisfied	2_Regular	0.292	0.199	0.292	0.217
1_Unsatisfied	4_Management	0.305	0.014	0.614	0.067
1_Unsatisfied	3_Senior	0.304	0.044	0.513	0.139
0_Unknown	1_Junior	0.301	0.397	0.106	0.196
0_Unknown	2_Regular	0.303	0.193	0.3	0.204
0_Unknown	4_Management	0.305	0.012	0.613	0.07
0_Unknown	3_Senior	0.301	0.038	0.535	0.126
'3_Highly satisfied'	1_Junior	0.306	0.411	0.1	0.184
'3_Highly satisfied'	2_Regular	0.303	0.187	0.305	0.204
'3_Highly satisfied'	4_Management	0.312	0.012	0.608	0.067
'3_Highly satisfied'	3_Senior	0.299	0.042	0.526	0.134
2_Satisfied	1_Junior	0.299	0.404	0.105	0.191
2_Satisfied	2_Regular	0.299	0.198	0.294	0.209
2_Satisfied	4_Management	0.306	0.016	0.613	0.065
2_Satisfied	3_Senior	0.291	0.043	0.536	0.13

From the above table, we see that in case the service requester (person who wants service) is a senior level or management level requester, the case is given higher priority while junior level person is given low priority, regular level requester is given low or medium priority. Thus, there is a bias in assigning case priority based on requester seniority.

Probability Distribution Table For 'Case Severity'

'Service Satisfaction'	'Case Type'	2_Normal	1_Minor	3_Major	4_Critical	0_Unclassified
1_Unsatisfied	Issue	0.858	0.053	0.051	0.025	0.013
1_Unsatisfied	Request	0.939	0.024	0.029	0.006	0.002
0_Unknown	Issue	0.853	0.037	0.07	0.03	0.009
0_Unknown	Request	0.927	0.018	0.044	0.009	0.002
'3_Highly satisfied'	Issue	0.834	0.032	0.091	0.035	0.008
'3_Highly satisfied'	Request	0.92	0.013	0.055	0.012	0.001
2_Satisfied	Issue	0.854	0.037	0.065	0.034	0.01
2_Satisfied	Request	0.935	0.017	0.038	0.008	0.002

We see almost all cases are issued Normal case severity. So, it would be feasible to remove the case severity attribute since almost all cases are classified as normal.

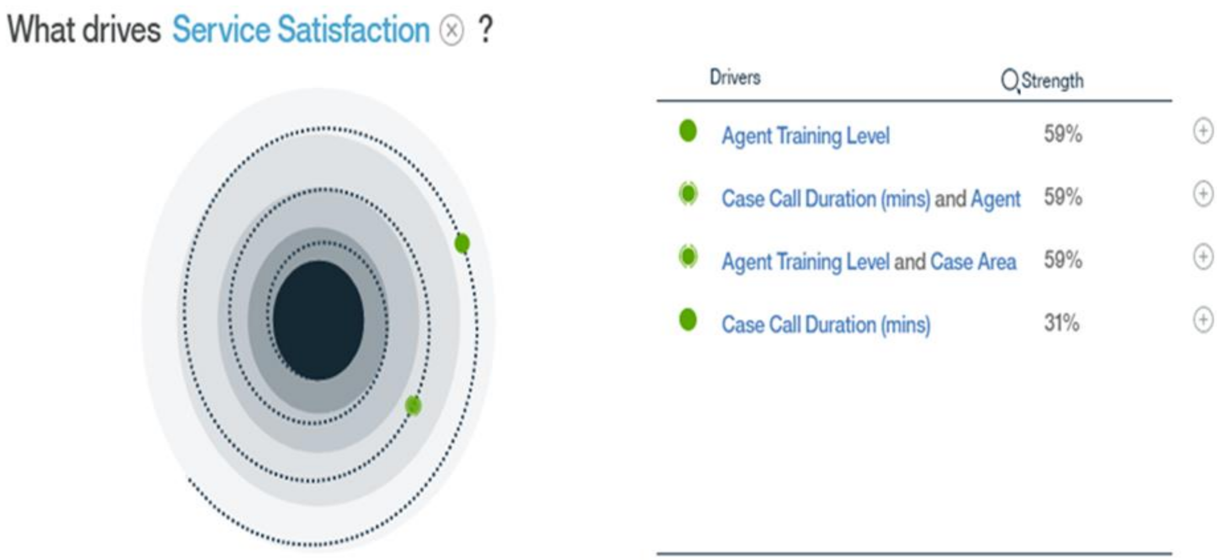
Cluster Analysis:

We do cluster analysis of dataset. We use Simple K-Means clustering method, set the number of clusters to 4 and we get the following result:

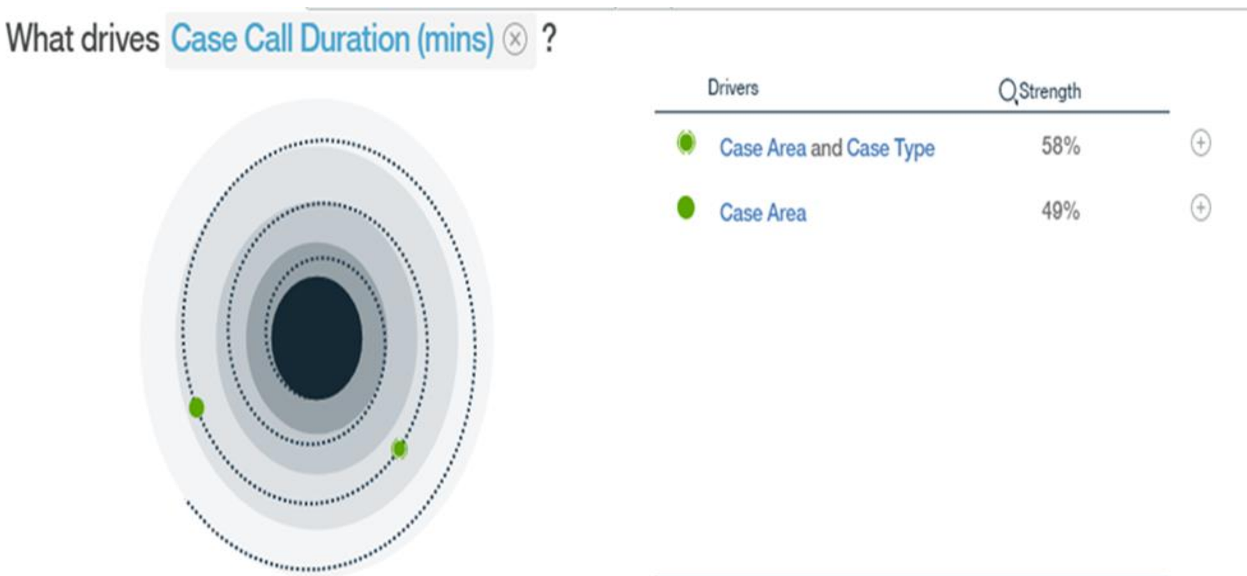
```
Number of iterations: 4
Within cluster sum of squared errors: 251961.94091364555

Initial starting points (random):

Cluster 0: Issue,Software,2_Normal,3_Senior,3_High,'1_Minimal training',2,1_Unsatisfied
Cluster 1: Request,Systems,2_Normal,2_Regular,3_High,'0_No training',13,1_Unsatisfied
Cluster 2: Request,Access/Login,2_Normal,2_Regular,2_Medium,'2_Sufficient training',0,2_Satisfied
Cluster 3: Request,Systems,3_Major,1_Junior,2_Medium,'0_No training',17,0_Unknown

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                          Full Data               0                 1                     2                 3
                                   (100000.0)        (16535.0)         (25397.0)             (41611.0)         (16457.0)
==========================================================================================================================
Case Type                             Request            Issue           Request               Request           Request
Case Area                             Systems         Software           Systems          Access/Login           Systems
Case Severity                        2_Normal         2_Normal          2_Normal              2_Normal          2_Normal
Requester Seniority                 2_Regular         3_Senior         2_Regular             2_Regular          1_Junior
Case Priority                          3_High           3_High            3_High           0_Unassigned          2_Medium
Agent Training Level    2_Sufficient training 1_Minimal training    0_No training 2_Sufficient training     0_No training
Case Call Duration (mins)              6.8428           4.4907            9.1009                4.1088           12.6343
Service Satisfaction                0_Unknown        0_Unknown    1_Unsatisfied     3_Highly satisfied         0_Unknown


Time taken to build model (full training data) : 0.34 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       16535 ( 17%)
1       25397 ( 25%)
2       41611 ( 42%)
3       16457 ( 16%)
```

The model generated by the cluster analysis correctly classifies highly satisfied and sufficient training together, Unsatisfied and no training together. Also, case severity is Normal in all clusters, so considering case severity as normal is common. So, case severity should be removed from dataset, just like we explained in classification. The cluster analysis is finding it difficult to cluster service satisfaction as 2_satisfied even if we increase the number of clusters to at least 5.

We then upload the dataset to IBM Watson so that we can see what conclusion can we derive after IBM Watson analyses the dataset.



The above figure shows the drivers for service satisfaction. We see that agent training level, call duration and case area are the most important drivers for service satisfaction. Thus, training for service agents need to be based on these 3 main factors. The accuracy of this dataset in predicting service satisfaction using Weka and IBM Watson is almost the same-around 59%.



From the above figure, we see that the main drivers of case call duration are case area and case type.

Dataset 3:

Sales Win-Loss Dataset:

This dataset provides information on sales result based on profits and losses.

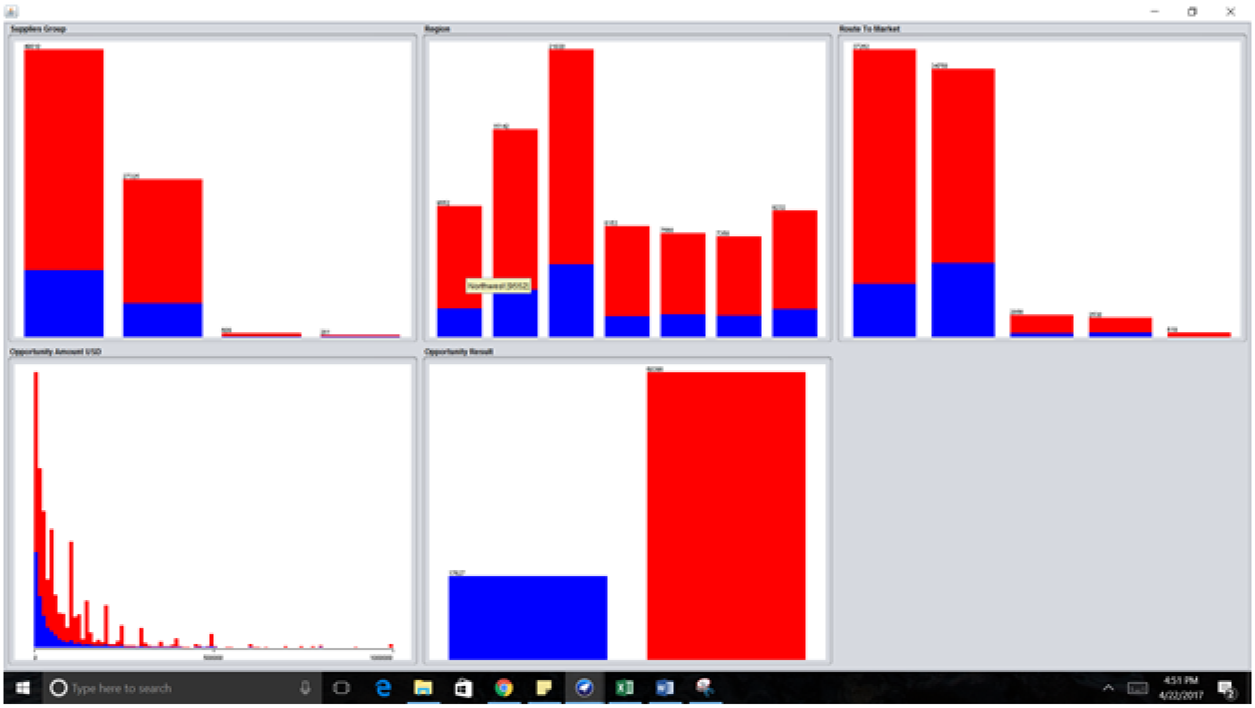
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Opportuni	Supplies Su	Supplies Gı	Region	Route To M	Elapsed Da	Opportuni	Sales Stage	Total Days	Total Days	Opportuni	Client Size	Client Size	Revenue F	Competito	Ratio Days	Ratio Days	Ratio Days	Deal Size Category	
1641984	Exterior Ac	Car Access	Northwest	Fields Sale	76	Won	13	104	101	0	5	5	0	Unknown	0.69636	0.113985	0.154215	1	
1658010	Exterior Ac	Car Access	Pacific	Reseller	63	Loss	2	163	163	0	3	5	0	Unknown	0	1	0	1	
1674737	Motorcycl	Performan	Pacific	Reseller	24	Won	7	82	82	7750	1	1	0	Unknown	1	0	0	1	
1675224	Shelters &	Performan	Midwest	Reseller	16	Loss	5	124	124	0	1	1	0	Known	1	0	0	1	
1689785	Exterior Ac	Car Access	Pacific	Reseller	69	Loss	11	91	13	69756	1	1	0	Unknown	0	0.141125	0	4	
1692390	Shelters &	Performan	Pacific	Reseller	89	Loss	3	114	0	232522	5	1	0	Unknown	0	0.000877	0	5	
1935837	Garage & (	Car Access	Pacific	Fields Sale	111	Won	12	112	112	20001	4	5	0	Unknown	0.308863	0.568487	0.12265	2	
1952571	Exterior Ac	Car Access	Pacific	Fields Sale	82	Loss	6	70	70	450000	1	1	0	Known	0.26361	0.73639	0	6	
1999486	Batteries &	Car Access	Northwest	Fields Sale	68	Loss	8	156	156	250000	1	5	0	None	0	0.562821	0.437179	6	
2052337	Exterior Ac	Car Access	Pacific	Reseller	18	Loss	7	50	50	55003	1	1	0	Unknown	0	0.585317	0.414683	4	
2100568	Exterior Ac	Car Access	Northwest	Fields Sale	76	Loss	8	165	165	0	1	2	0	Unknown	0.417729	0.23558	0.346691	1	
2190367	Garage & (	Car Access	Midwest	Fields Sale	87	Loss	5	142	142	400000	5	5	0	Known	0.015482	0.370162	0.614356	6	
2217068	Performan	Performan	Midwest	Reseller	35	Loss	6	31	31	10000	2	1	0	Unknown	0	0.167213	0.832787	2	
2223143	Exterior Ac	Car Access	Pacific	Reseller	16	Loss	5	208	208	232522	1	1	0	Unknown	0.946076	0.053924	0	5	
2228661	Batteries &	Car Access	Midwest	Fields Sale	81	Loss	10	138	138	200000	4	5	4	Known	0	0.730044	0.269956	5	
2228983	Batteries &	Car Access	Northwest	Fields Sale	79	Won	5	32	32	0	5	1	0	Known	0.024845	0.456522	0.518634	1	
2263363	Towing & I	Car Access	Midwest	Reseller	83	Loss	13	130	130	60000	4	3	0	Unknown	0.12182	0.558982	0.319198	4	
2277276	Garage & (	Car Access	Pacific	Fields Sale	65	Loss	17	150	150	250009	5	5	0	Known	0.068182	0.625	0.306818	6	
2280685	Shelters &	Performan	Northwest	Fields Sale	91	Loss	6	103	103	500000	1	3	0	Known	0	1	0	7	
2284303	Shelters &	Performan	Northwest	Fields Sale	65	Loss	13	125	125	100000	1	5	0	None	0.029695	0.128411	0.841894	5	
2289905	Shelters &	Performan	Pacific	Fields Sale	89	Loss	7	60	60	150000	5	4	0	Unknown	0.313433	0.686567	0	5	
2292996	Shelters &	Performan	Midwest	Reseller	62	Loss	14	88	87	210000	5	3	0	Unknown	0.041096	0.8379	0.11758	5	
2296022	Exterior Ac	Car Access	Midwest	Reseller	16	Loss	8	169	169	3000	3	2	0	Unknown	0.189125	0.810875	0	1	
2315483	Motorcycl	Performan	Midwest	Fields Sale	83	Loss	7	90	48	80000	3	5	0	Unknown	0	0.403352	0.134078	4	
2315706	Motorcycl	Performan	Northwest	Reseller	73	Won	9	127	127	40721	1	1	1	Unknown	0.403467	0.43814	0.158392	3	
2315732	Motorcycl	Performan	Midwest	Reseller	76	Loss	8	133	133	64000	1	4	0	Unknown	0.141679	0.344828	0.513493	4	
2345802	Motorcycl	Performan	Midwest	Fields Sale	86	Loss	2	119	119	51000	1	3	0	Unknown	1	0	0	4	

WA_Fn-UseC_-Sales-Win-Loss

As done in the earlier datasets, we considered Supplies Group, Region, Route to Market and Opportunity revenue to predict the loss and win result of clients.

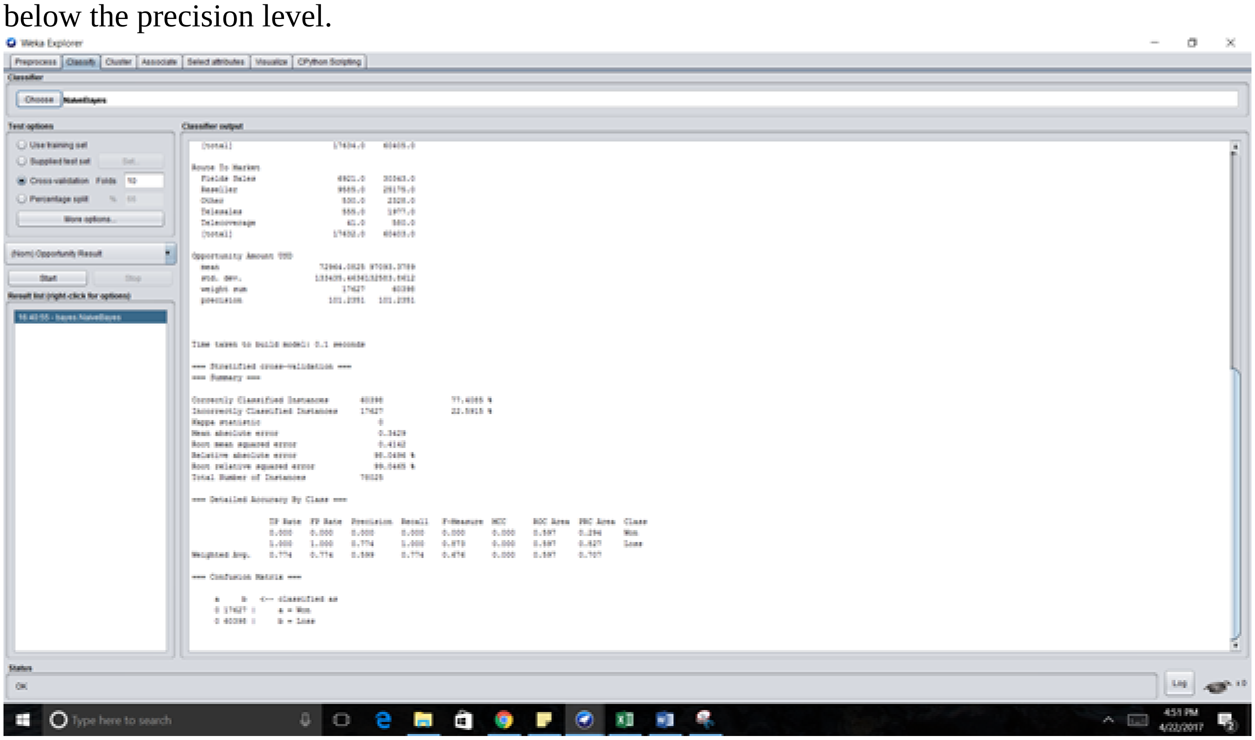
1 Supplies Group
2 Region
3 Route To Market
4 Opportunity Amount USD
5 Opportunity Result

The below graphs provide various descriptions about the statistics of the variables chosen to predict Opportunity result. The below graphs provide information regarding maximum, minimum, mean and standard deviation.



NaiveBayes

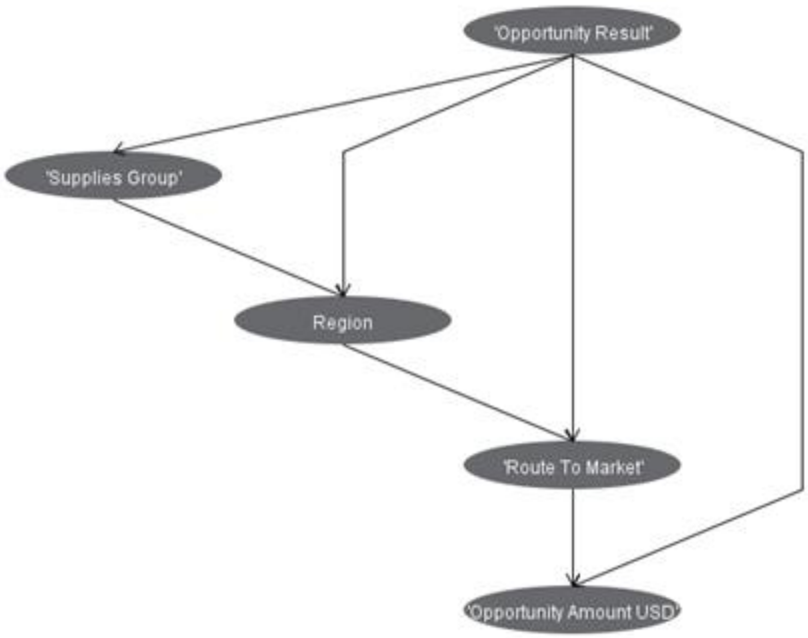
We performed Naive Bayes and found that the accuracy to be nearly 80% with the recall level below the precision level.



BayesNet:

We performed a BayesNet algorithm along with cross validation to get the visualized graph for Opportunity result. We can infer that there were more number of Losses than Wins from the confusion matrix.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       61418               78.7158 %
Incorrectly Classified Instances     16607               21.2842 %
Kappa statistic                          0.3002
Mean absolute error                      0.2803
Root mean squared error                  0.3772
Relative absolute error                 80.1276 %
Root relative squared error             90.205  %
Total Number of Instances            78025

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.345    0.084    0.546      0.345    0.423      0.312    0.785     0.483     Won
                 0.916    0.655    0.827      0.916    0.870      0.312    0.785     0.919     Loss
Weighted Avg.    0.787    0.526    0.764      0.787    0.769      0.312    0.785     0.821

=== Confusion Matrix ===

     a     b   <-- classified as
  6079 11548 |     a = Won
  5059 55339 |     b = Loss
```



The above BayesNet graph is obtained by providing maximum number of parents as 2. The below probability distribution table provides information about how Opportunity result is affected by both region and Route to Market. The reseller working in the Northwest and has an opportunity result of Won has a probability of 0.547.

Probability Distribution Table For 'Route To Market'

'Opportunity Result'	Region	'Fields Sales'	Reseller	Other	Telesales	Telecoverage
Won	Northwest	0.414	0.547	0.036	0.003	0.001
Won	Pacific	0.444	0.403	0.08	0.067	0.005
Won	Midwest	0.331	0.615	0.017	0.034	0.002
Won	Southwest	0.417	0.541	0.011	0.029	0.001
Won	Mid-Atlantic	0.335	0.642	0.016	0.005	0.002
Won	Northeast	0.414	0.556	0.019	0.009	0.001
Won	Southeast	0.453	0.502	0.009	0.034	0.002
Loss	Northwest	0.507	0.453	0.027	0.006	0.007
Loss	Pacific	0.461	0.317	0.127	0.072	0.023
Loss	Midwest	0.53	0.415	0.021	0.027	0.006
Loss	Southwest	0.494	0.442	0.011	0.051	0.002
Loss	Mid-Atlantic	0.471	0.494	0.014	0.02	0.001
Loss	Northeast	0.534	0.437	0.011	0.013	0.005
Loss	Southeast	0.511	0.441	0.013	0.019	0.016

From the above probability distribution function, we find that we need to focus our attention on field sales and resellers, and improve our relationship with them since they are giving us better market penetration. We need to train our field sales agents since the probability of loss is more for field sales than resellers and probability of a win is more for resellers than field sales. Field sales and resellers are also important for future sales since the opportunity cost is dependent on Route to Market and opportunity result. If we lose focus on field sales and resellers, we may have more lost opportunities in the future.

Probability Distribution Table For Region

'Opportunity Result'	'Supplies Group'	Northwest	Pacific	Midwest	Southwest	Mid-Atlantic	Northeast	Southeast
Won	'Car Accessories'	0.114	0.187	0.318	0.085	0.104	0.08	0.112
Won	'Performance & Non-auto'	0.125	0.216	0.266	0.093	0.073	0.108	0.119
Won	'Tires & Wheels'	0.071	0.071	0.561	0.045	0.084	0.071	0.097
Won	'Car Electronics'	0.084	0.084	0.432	0.123	0.084	0.071	0.123
Loss	'Car Accessories'	0.127	0.199	0.255	0.111	0.107	0.09	0.112
Loss	'Performance & Non-auto'	0.12	0.186	0.264	0.107	0.083	0.108	0.133
Loss	'Tires & Wheels'	0.077	0.135	0.396	0.086	0.11	0.083	0.112
Loss	'Car Electronics'	0.116	0.23	0.378	0.097	0.05	0.055	0.074

From the above table, we find that the Midwest region and Pacific region has a better opportunity as compared to other regions since we get better probabilities of a win, so we need to increase market penetration in Midwest and the Pacific regions as it may increase probability of a win. So, we need to focus more attention on selling tires and wheels, car electronics in Midwest. In the Pacific region, we need to pay more attention on selling car accessories and performance and Non-auto accessories.

Clustering:

Clustering is an unsupervised machine learning technique which is used for grouping items by probability which are like each other using Euclidean distance with no prior knowledge of classes.

Here we use EM for clustering. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify apriori how many clusters to generate. The below cluster information is provided in the below screenshot. We can infer that the Win ratio is always lesser than the Loss ratio in every cluster.

EM
==

Number of clusters selected by cross validation: 6
Number of iterations performed: 100

Attribute	Cluster 0 (0.33)	1 (0.04)	2 (0.23)	3 (0.23)	4 (0.14)	5 (0.03)
Supplies Group						
Car Accessories	16347.3076	3015.513	11205.1842	12584.4485	6586.6572	76.8894
Performance & Non-auto	9234.8477	71.3313	6652.243	5029.0024	4445.6488	1897.9268
Tires & Wheels	141.5483	13.4085	216.3799	70.7152	148.9356	24.0125
Car Electronics	74.1473	21.9896	40.0057	101.7752	40.9229	8.1592
[total]	25797.8509	3122.2424	18113.8129	17785.9413	11222.1646	2006.988
Region						
Northwest	3287.0042	408.7274	2339.1339	1772.8315	1483.3939	266.9089
Pacific	3733.2999	878.9433	3755.8422	3865.1016	2448.2332	466.5799
Midwest	6818.2562	680.0763	4814.3856	5567.8413	2701.3236	444.1169
Southwest	2824.3877	325.8241	2166.3972	1566.9985	1110.4251	164.9674
Mid-Atlantic	2991.7304	192.2117	1475.0074	1710.0783	1037.6311	167.3411
Northeast	2740.751	237.1764	1845.3179	1205.7769	1098.8906	236.0873
Southeast	3405.4214	402.2832	1720.7287	2100.3131	1345.267	263.9866
[total]	25800.8509	3125.2424	18116.8129	17788.9413	11225.1646	2009.988
Route To Market						
Fields Sales	7449.9449	2211.2556	11449.5809	6381.719	8323.9398	1451.5598
Reseller	16907.1681	677.7536	5147.3617	9608.3083	2022.1796	401.2287
Other	477.3522	175.8466	830.3187	602.329	634.5558	141.5977
Telesales	747.6505	38.4201	517.6506	1056.9422	166.9637	8.3729
Telecoverage	216.7352	19.9665	169.901	137.6428	75.5257	5.2289
[total]	25798.8509	3123.2424	18114.8129	17786.9413	11223.1646	2007.988
Opportunity Amount USD						
mean	31696.5927	464903.5689	92702.9515	7583.4203	189080.627	473252.6799
std. dev.	15741.7723	217061.1337	31096.1783	5819.6678	72549.3501	225711.9793
Opportunity Result						
Won	5808.3342	771.8642	2446.153	6744.8706	1480.4523	381.3257
Loss	19987.5166	2348.3782	15665.6599	11039.0707	9739.7123	1623.6623
[total]	25795.8509	3120.2424	18111.8129	17783.9413	11220.1646	2004.988

The below screenshot provides the percentage of clustered instances in each of the cluster (0-5).

Clustered Instances

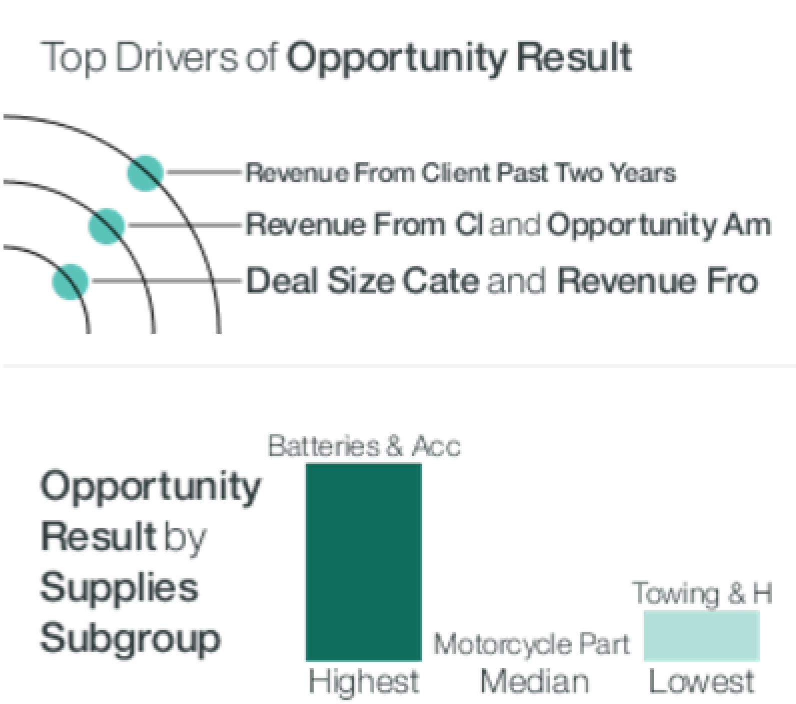
0	26093	(33%)
1	2452	(3%)
2	17356	(22%)
3	19932	(26%)
4	10578	(14%)
5	1614	(2%)

IBM Watson Analytics:

We consider all the parameters in the dataset. Opportunity Number, Supplies Subgroup, Supplies Group, Region, Route To Market, Elapsed Days In Sales Stage, Opportunity Result, Sales Stage Change Count, Total Days Identified Through Closing, Total Days Identified Through Qualified, Opportunity Amount USD, Client Size By Revenue, Client Size By Employee Count, Revenue

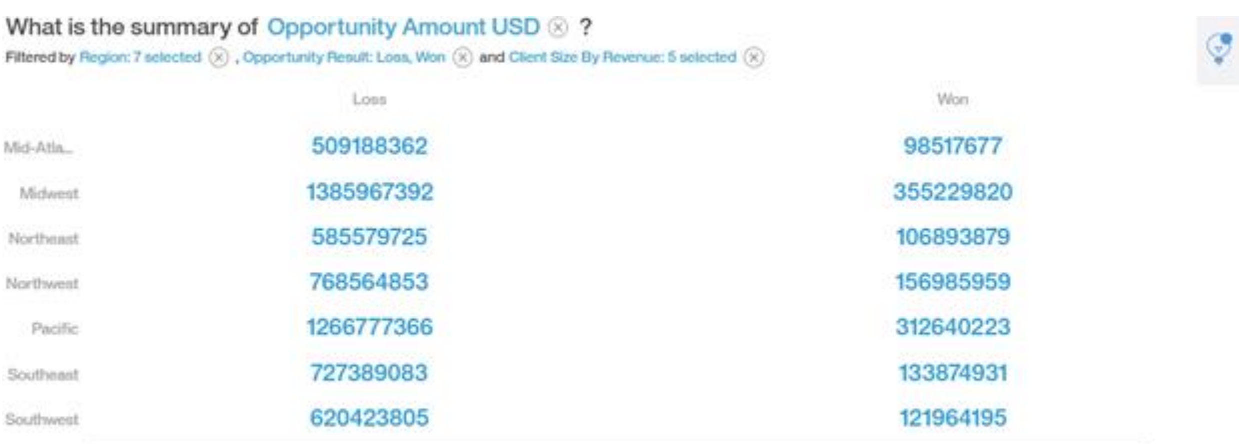
From Client Past Two Years, Competitor Type, Ratio Days Identified To Total Days, Ratio Days Validated To Total Days, Ratio Days Qualified To Total Days, Deal Size, Category

We used IBM Watson Analytics to predict the best driver for opportunity result.



The graph provides information regarding the four attributes which were considered for data analysis. We can observe the graph for Loss is more non-linear than Won. This result utilizes client revenue of 2 years to predict the opportunity win-loss ratio.

We can also calculate summary statistics for any attribute in the dataset using IBM Watson Analytics.



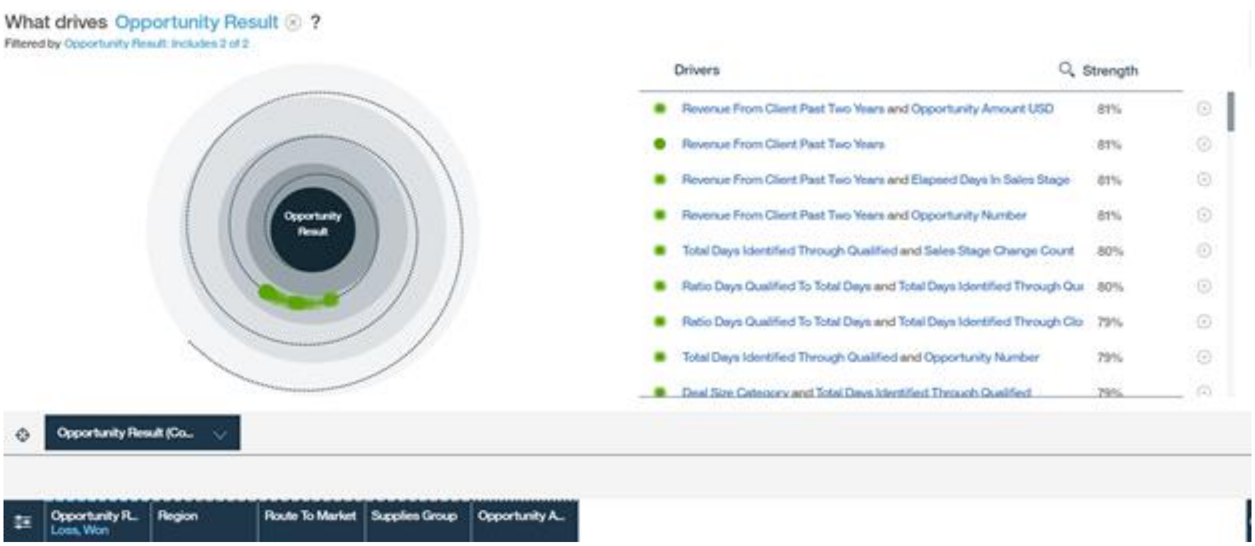
What is the summary of Opportunity Amount USD ?
Filtered by Region: 7 selected , Opportunity Result: Loss, Won and Client Size By Revenue: 5 selected

	Loss	Won
Mid-Atla...	509188362	98517677
Midwest	1385967392	355229820
Northeast	585579725	106893879
Northwest	768564853	156985959
Pacific	1266777366	312640223
Southeast	727389083	133874931
Southwest	620423805	121964195

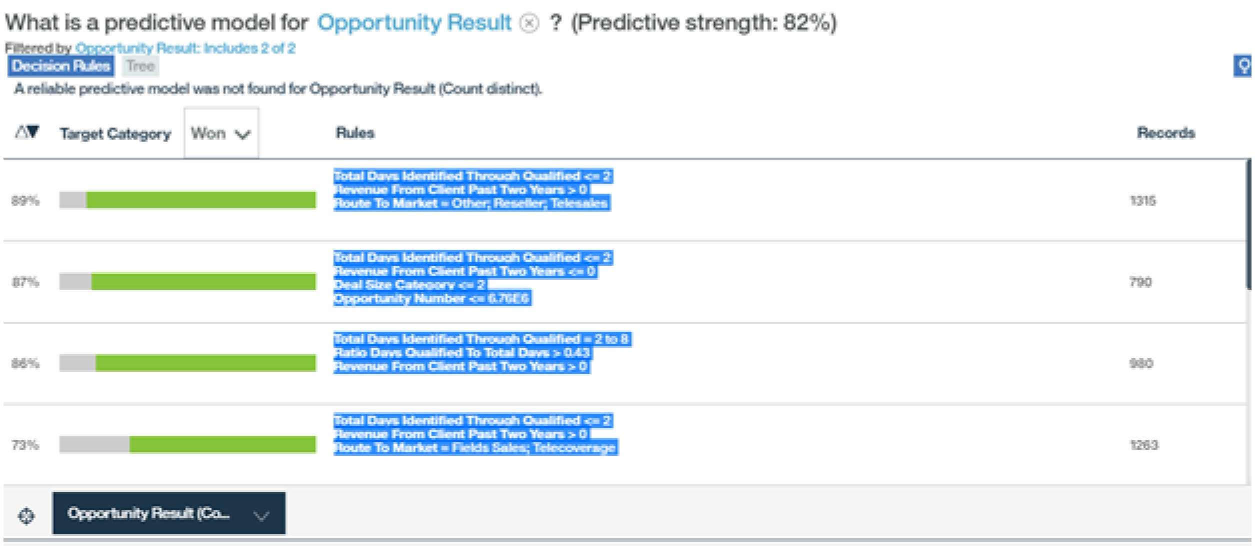
What is the summary of Revenue From Client Past Two Years ?
Filtered by Region: 7 selected and Client Size By Employee Count: 5 selected

	Mid-Atlantic	Midwest	Northeast	Northwest	Pacific	Southeast	Southwest
1	1934	5507	1481	2182	2885	2233	1728
2	68	220	30	100	159	127	87
3	118	347	97	174	144	122	58
4	135	743	99	160	145	160	36
5	348	1068	71	279	368	65	186

The spiral graph provided information regarding the top drivers in predicting Opportunity result. On observation of spiral graph, we can conclude that the closer the combination is to the core attribute, the more strength it must predict that core attribute. The best predictive model can also be calculated using IBM Watson Analytics.



Thus, we see that the two most important drivers for opportunity result is Revenue from client for past two years and opportunity amount in USD.



Comparing with Weka:

Similar analysis was done on Weka and we could replicate the same results obtained from IBM Watson Analytics. This provides the accuracy and usefulness of IBM Watson Analytics.

J48 algorithm:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       63743               81.6956 %
Incorrectly Classified Instances     14282               18.3044 %
Kappa statistic                          0.3712
Mean absolute error                      0.2713
Root mean squared error                  0.3695
Relative absolute error                 77.5748 %
Root relative squared error             88.3663 %
Total Number of Instances            78025

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.358    0.049    0.680      0.358    0.469      0.399    0.762     0.550     Won
                 0.951    0.642    0.835      0.951    0.889      0.399    0.762     0.889     Loss
Weighted Avg.    0.817    0.508    0.800      0.817    0.794      0.399    0.762     0.812

=== Confusion Matrix ===

     a     b   <-- classified as
  6313 11314 |     a = Won
  2968 57430 |     b = Loss
```

Dataset 4:

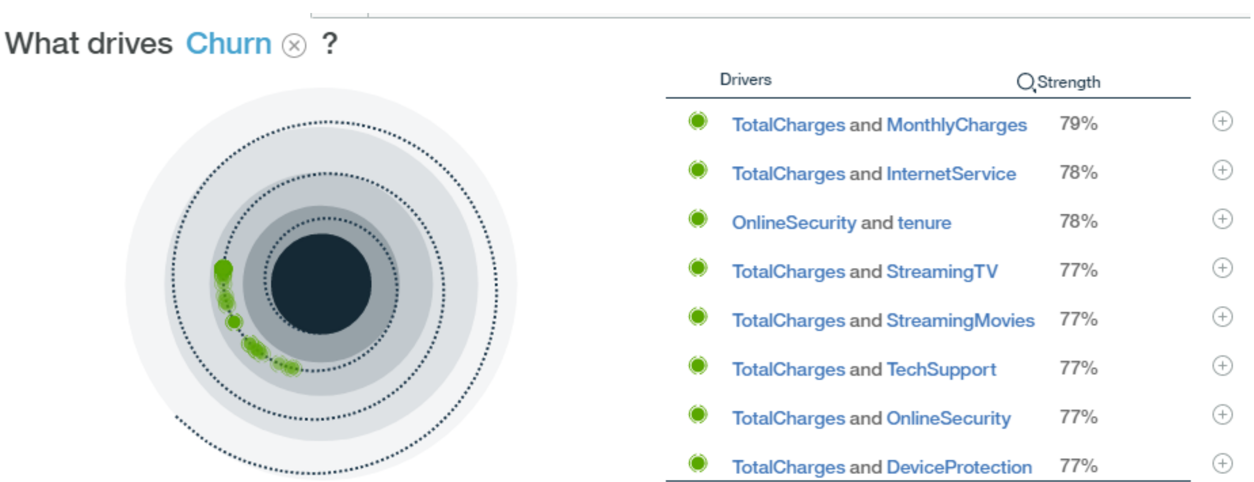
Churn dataset:

Here Churn is predicted based on Total Charges in the dataset for any department in a company.

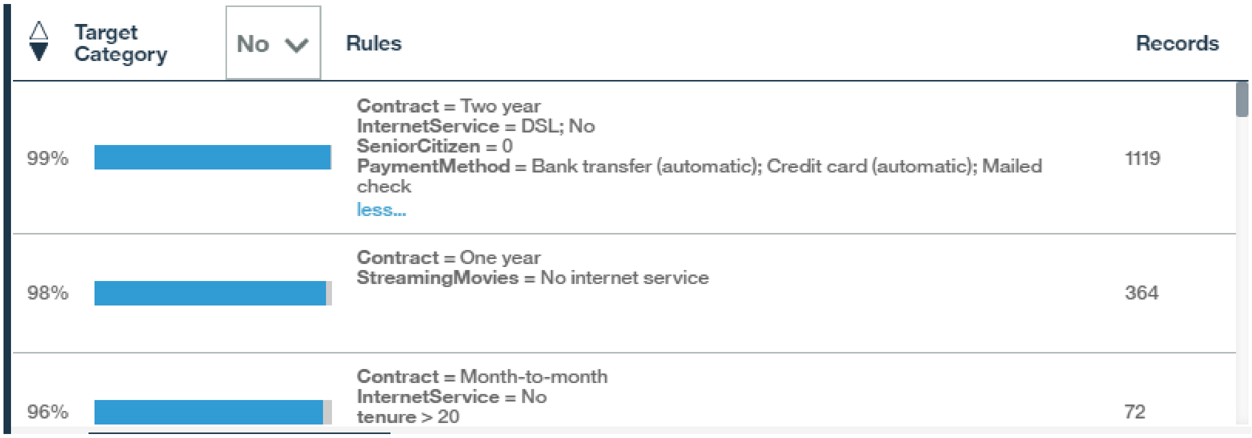
We consider all the parameters in the dataset- CustomerID, Gender, SeniorCitizen, Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,StreamingTV,StreamingMovies,Contract,PaperlessBilling, PaymentMethod,MonthlyCharges,TotalCharges and Churn.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
customerI	gender	SeniorCitiz	Partner	Dependent	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSecu	OnlineBacl	DeviceProt	TechSuppc	StreamingT	StreamingI	Contract	PaperlessB	PaymentM	MonthlyCh	TotalCharg	Churn
7590-VHVI	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No
5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No
3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed che	53.85	108.15	Yes
7795-CFO(	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
9237-HQIT	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes
9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes
1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit carc	89.1	1949.4	No
6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed che	29.75	301.9	No
7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes
6388-TAB(	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
9763-GRSI	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed che	49.95	587.45	No
7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit carc	18.95	326.8	No
8091-TTVA	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit carc	100.35	5681.1	No
0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No
3655-SNQ'	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit carc	113.25	7895.15	No
8191-XWS	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed che	20.65	1022.95	No
9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
4190-MFLI	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit carc	55.2	528.35	Yes
4183-MYFI	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No
8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic	39.65	39.65	Yes
1680-VDC\	Male	0	Yes	No	12	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Bank trans	19.8	202.25	No
1066-JKSG	Male	0	No	No	1	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Month-to-	No	Mailed che	20.15	20.15	Yes
3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit carc	59.9	3505.1	No
6322-HRPI	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-	No	Credit carc	59.6	2970.3	No
6865-JZNK	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Bank trans	55.3	1530.6	No
6467-CHFZ	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to-	Yes	Electronic	99.35	4749.15	Yes

As Churn is dependent on Total Charges, we would like to analyze the important factors which affect Total charges. Using IBM Watson Analytics, we can find that the top drivers for Total Charges are Monthly Charges and Tenure.

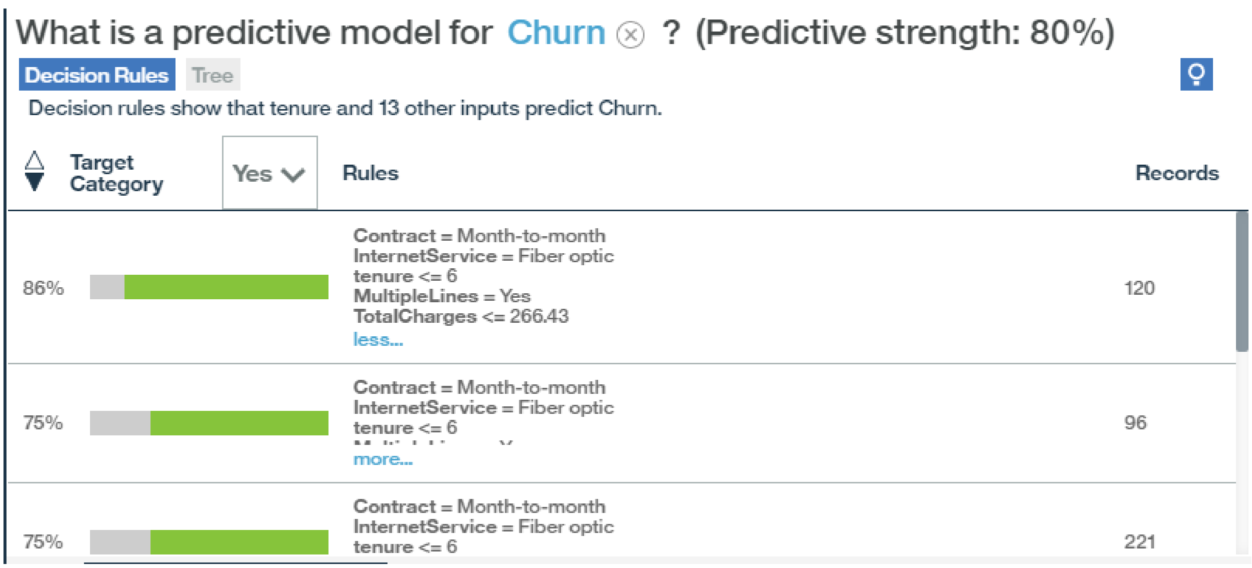


We see that customer churn is driven by Total Charges, monthly charges, internet service, online security and tenure (total amount of time a customer has subscribed to service).



Target Category: No	Rules	Records
99%	Contract = Two year InternetService = DSL; No SeniorCitizen = 0 PaymentMethod = Bank transfer (automatic); Credit card (automatic); Mailed check less...	1119
98%	Contract = One year StreamingMovies = No internet service	364
96%	Contract = Month-to-month InternetService = No tenure > 20	72

The above figure shows predictive model of no churn, which is desirable. The contract should be a 2-year contract, internet service = DSL, not a senior citizen and payment is through bank transfer, credit card and mailed check.



What is a predictive model for Churn ? (Predictive strength: 80%)

Decision Rules Tree

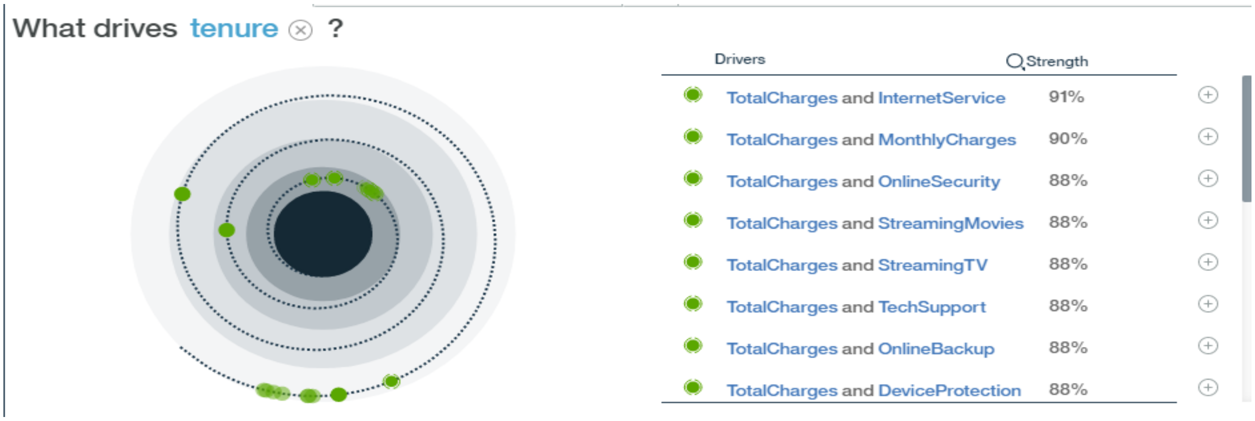
Decision rules show that tenure and 13 other inputs predict Churn.

Target Category: Yes	Rules	Records
86%	Contract = Month-to-month InternetService = Fiber optic tenure <= 6 MultipleLines = Yes TotalCharges <= 266.43 less...	120
75%	Contract = Month-to-month InternetService = Fiber optic tenure <= 6 more...	96
75%	Contract = Month-to-month InternetService = Fiber optic tenure <= 6	221

The above figure shows predictive model when churn happens, which is not desirable. The contract is month to month. Internet service is fiber optic, tenure is less than or equal to 6 months.

Tenure:

Tenure is important since we want to see the factors which help in keeping customers a subscription with the company for a longer time:



Thus, we see that total charges, monthly charges, internet service, online security are important factors for tenure.

What is a predictive model for tenure ? (Predictive strength: 93%)

Decision Rules | Tree

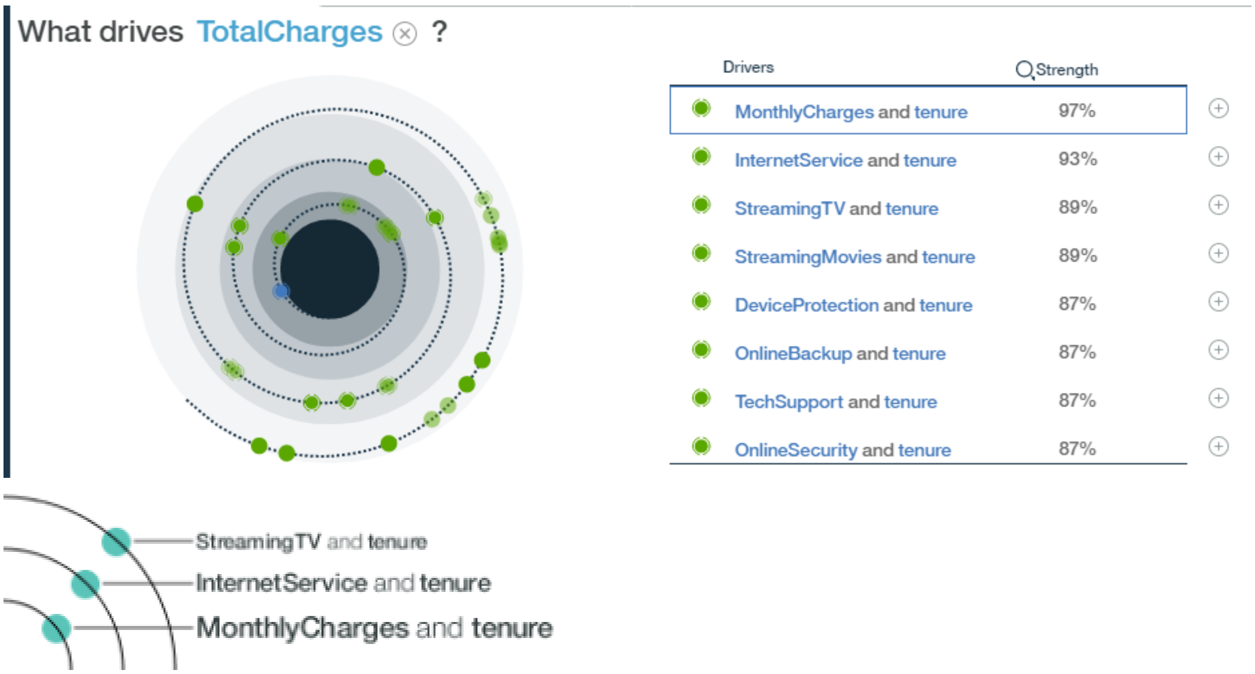
Decision rules show that MonthlyCharges and 14 other inputs predict tenure.

Predicted value	Rules	Records
69.48	TotalCharges > 4,476.85 Contract = Two year Partner = Yes OnlineBackup = Yes OnlineSecurity = Yes less...	324
67.68	TotalCharges > 4,476.85 Contract = Two year Partner = Yes OnlineBackup = Yes OnlineSecurity = No less...	106
66.77	TotalCharges > 4,476.85 Contract = Two year Partner = Yes	124

The above figure shows predictive model for tenure, contract should be a 2-year contract, customer should have partner, online backup and online security.

Total charges

Since in the analysis of tenure and customer churn, total charges were a significant driver of both attributes, we also analysis the drivers for total charges:

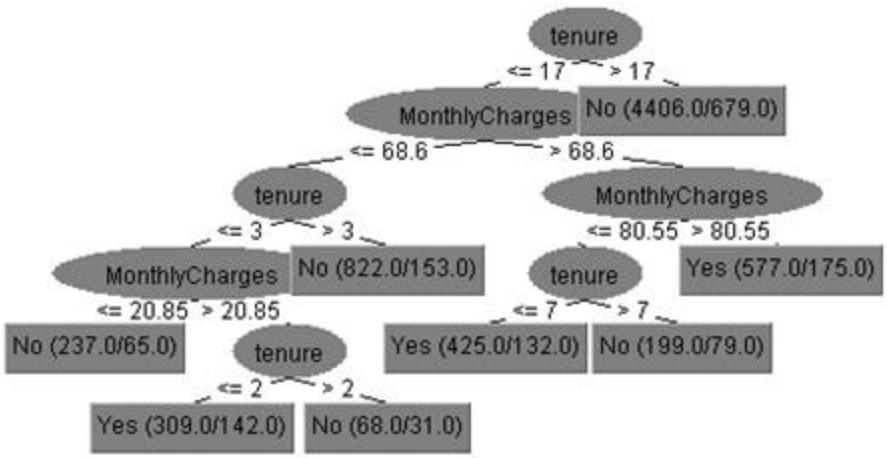


Thus, we see that tenure, monthly charges, internet service, streaming services (TV and Movies), device protection are the most important factors that contribute to total charges. This means that these factors can help boost revenue.

As a result, we have taken Tenure and Monthly Charges along with Total Charges to predict Churn for a company.

No.	Name
1	tenure
2	MonthlyCharges
3	TotalCharges
4	Churn

The below graph provides information about the visual graph using classifier NaiveBayes for tenure.

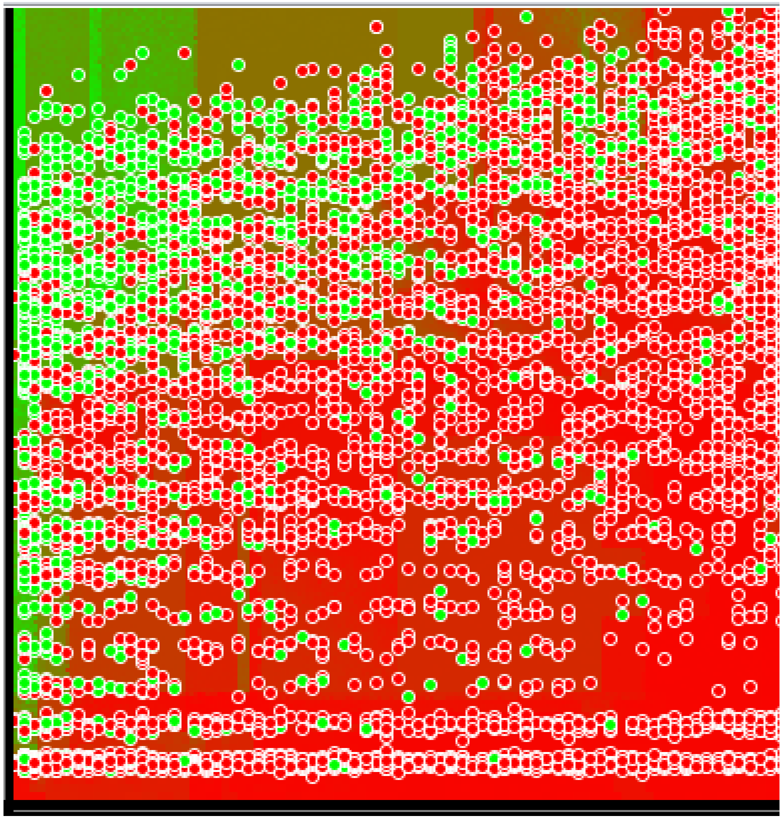


Boundary Visualizer using Adaboost M1 with J48:

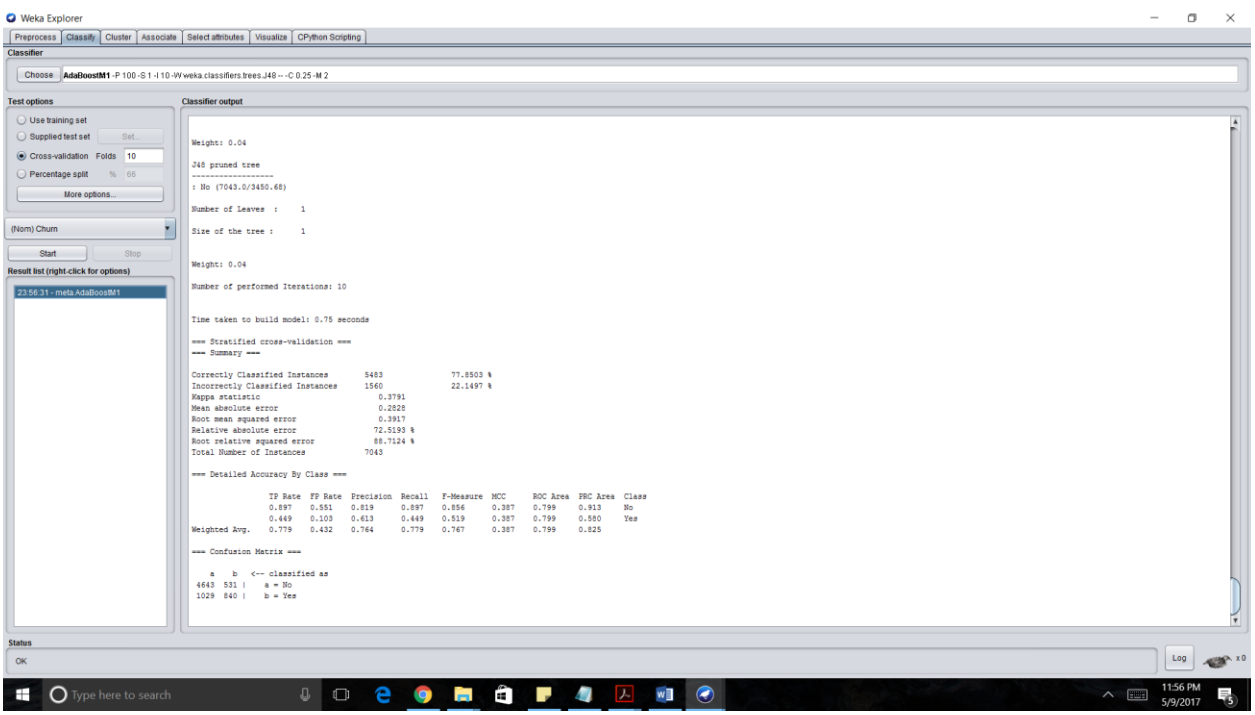
We wanted to analyze the dataset using Boundary Visualizer. For this purpose, we used Adaboost meta-learning technique using J48 algorithm. Here the class attribute is Churn. Adaboost is a boosting technique used along with another classifier algorithm to further refine the model to provide more classification of instances.
In the two-class model, for the weight αtαt, we use:

αt=ln1−ϵtϵt.

The below visualization provides an in-depth view of the data points and classification method. The green dots refer to Yes and red dots refer to No. The greener the area the easier it was to classify the data points by the respective classifier. We can observe that there is a strong mix of Yes and No even in darker areas bringing the accuracy down for predicting Churn.

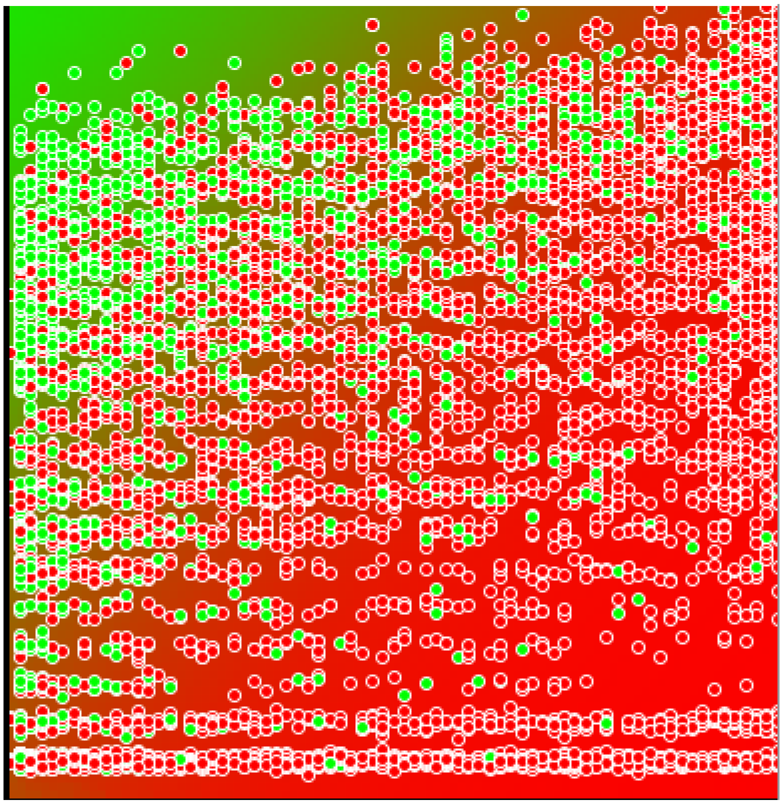


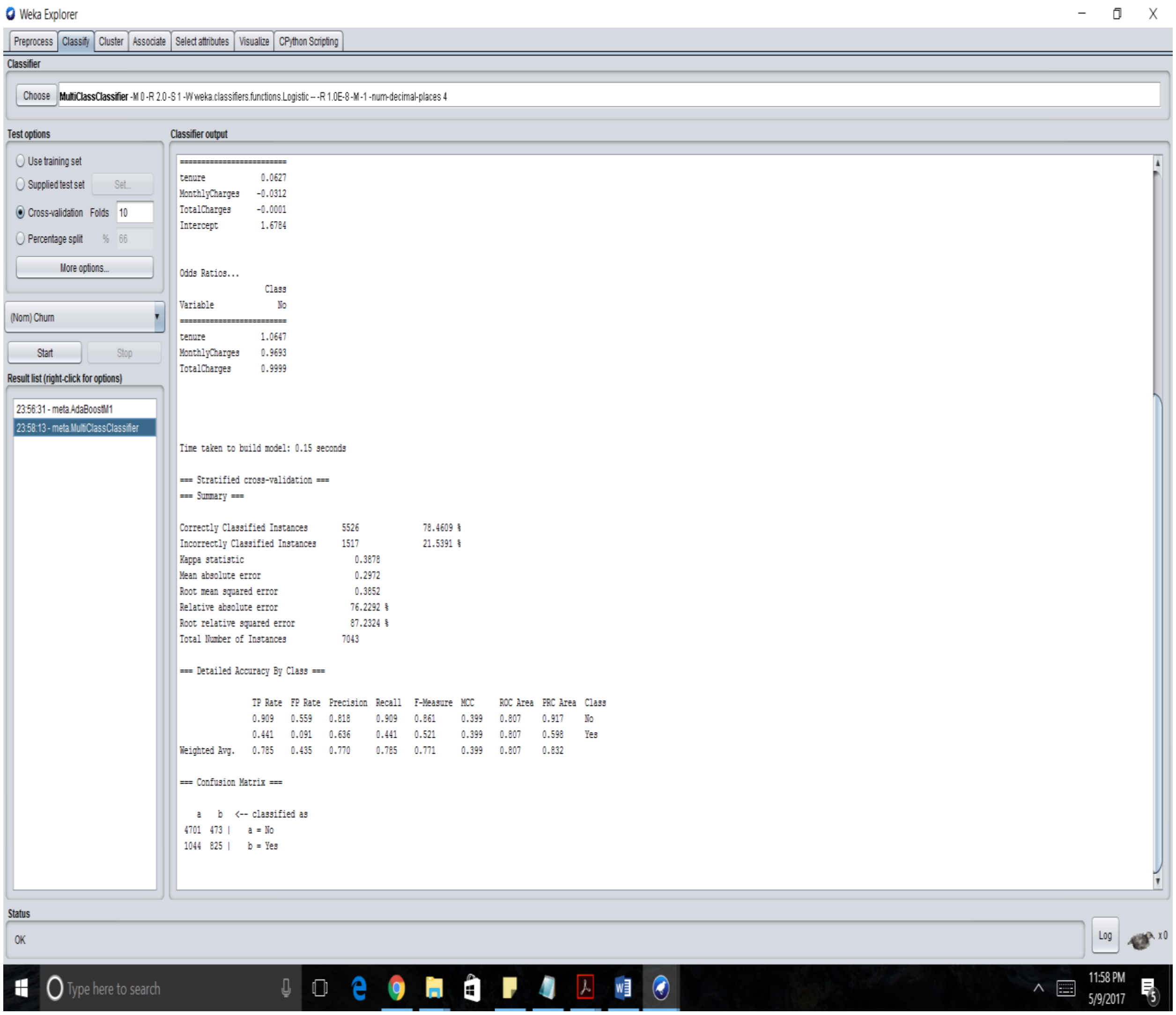
The accuracy was 77%



Boundary Visualizer using Multi-ClassClassifier using Logistic

We tried to use Multi-ClassClassifier using Logistic regression and found out that the visualization was smoother and the accuracy improved by a very small margin. Logistic regression seems to classify the data points much better than J48.





Recommendations:

A) Recommendations to reduce attrition rate and improve employee job involvement:

1) Managers need to be trained in managing how to deal with the employees of their department.
2) The HR can use the following formula to predict the number of years a person will work in the company:

   (-0.0323 * Age) + (-0.2923 * NumCompaniesWorked)+

   (-0.0367 * PercentSalaryHike) + (0.2716 * TotalWorkingYears) +

   (0.4657 * YearsInCurrentRole) + (0.3187 * YearsSinceLastPromotion) +

(0.5578 * YearsWithCurrManager) + 1.5157

The HR should also look at the attributes: year with current manager, total number of working years and the number of companies the employee has changed to predict if the employee will work for company for at least 7 years.
Since every employee who leaves company is an expense in terms of money, experience and knowledge, we need to try to ensure that an employee stays within company for at least 7 years. An employee redressal needs to be setup in case an employee is dissatisfied with job and recommendations of each employee needs to be considered and acknowledged. In case employees complain of any issue in their departments, we must ensure anonymity of the complainants and solve the complains.

3) To increase job satisfaction, the HR need to consider the employee’s Job Role, years at company, age, years in current role, total working years and years with current manager
4) For employees to stay closer to office, housing allowance needs to be increased or housing complexes near to office needs to be built.
5) Travelling for business work must be restricted and not dependent on a few employees, other employees also should go on business travel.
6) People in the age groups 26 to 42 years tend to have better job involvement, more salary hikes and better work-life balance. So, we need to improve the job involvement of employees in other age groups also.
7) We should consider steps as to how to increase job involvement of employees after time since employee gets last promotion increases.
8) Training sessions for all employees need to be increased. Employees with performance rating less than 3 should be put through additional training so that the employee’s performance rating will increase. Also, employees should be encouraged to reduce or avoid overtime by completing the work during office hours.
9) Employees should be encouraged to buy stock options and should be encouraged on taking challenging projects.

B) Recommendations for the service agents to improve service satisfaction.

1) Invest on training of employees.
2) Train employees more on hardware cases since these cases seem more complex to solve, After hardware, the order of difficulty in solving the cases (hardest to the easiest) is software, systems and login.
3) There seems to be a bias in assigning case priority based on requester seniority which needs to be addressed. Case priority should depend on how much the case may affect the organization than on requester seniority.
4) Almost all cases are assigned as Normal case severity. So, it is feasible to remove this attribute since it does not seem to have an impact on service satisfaction.
5) Agent training level, call duration and case area are the most important drivers for service satisfaction.
6) Case area and case type the main drivers of case call duration.

C) Recommendations to increase the company's sales potential:

1) Use the resellers and field agents to increase market penetration.
2) Training of field agents need to be improved and increased and should be made on par with resellers.
3) Midwest and Pacific regions of the US are more likely to have better opportunity results. The company needs to focus more attention and increase market penetration of products in these regions.
4) To focus more attention on selling tires and wheels, car electronics in Midwest. In the Pacific region, we need to pay more attention on selling car accessories and performance and Non-auto accessories.
5) We need to improve relations with clients from whom we have earned more income in the past 2 years, and look at the opportunity costs if we want to increase chances of a win.
6) The two most important drivers for opportunity result is Revenue from client for past two years and opportunity amount in USD.

D) Recommendations to reduce customer churn and increase tenure:

1) The contract time should be increased to 2 years, so customers need to be encouraged to sign a 2-year contract through incentives.
2) DSL internet service needs to be promoted and we need to study the reasons why customers who have fiber optic connections are leaving the service.
3) Total charges are an important factor in customer churn and tenure. But we need to maintain the price point since we do not see evidence that the cost is high or less in the analysis.
4) We need to ensure that a new customer stays in the company for at least 6 months or else the chances that the customer leaves the service increases.
5) We see that tenure, monthly charges, internet service, streaming services (TV and Movies), device protection are the most important factors that contribute to total charges. These sources are important to improve revenue. Thus, we must ensure that customer's tenure is longer, and we need to promote internet service, streaming services and device protection services to customers to increase revenue.
6) To increase tenure, contract should be a 2-year contract, we need to target customers who have a partner, and are interested in the online backup, online security services and internet services. Tenure predictive model shows that online backup and online security are potential revenue earners, so we need to first enhance and then promote these two services.
7) We need to promote payment through bank transfer, credit card and mailed check to reduce customer. Through the predictive model of customer churn, we find that the services are more suitable for younger customers. This is because younger customers understand technology well and have a need for the services.

Conclusion:

Through a combination of using the strengths of Weka and IBM Watson, we can do a comprehensive analysis of a dataset and can present the findings to the director. IBM Watson helped us with the visualizations. We did analysis on Weka to check the accuracy of the prediction analysis and to obtain inferences by using BayesNet. The visualization of BayesNet analysis in Weka along with the probability distribution tables (P.D.T.) helped us in understanding the dependencies of the attributes. The BayesNet P.D.T. gave us important inferences which were surprising to us. This project taught us valuable insights on the how data mining and visualization are important tools for improving various aspects of industry. Also, we learned how we can interpret the results of the data analysis and present our findings to other people in a human readable form. This project taught us various analytical skills which we can use in the future.