Assignment A2: Text + Clustering + Estimation			
Student Name	**Jose Arturo Gil Alonso**	**Student No**	*218659676*
My other group members		**A2 Group No**	
Team Names	**Omar Malaeb**	**Student Nos**	*218453505*

	Exceptional	Meets expectations	Issues noted	Improve	Unacceptable
Exec Report					
Explore Attributes					
Discover Relationships					
Create Models					
Evaluate & Improve					
Provide Solution					
Research & Extend					
Brief Comments					**Total**

Executive summary (one page)

Expectation

This report aims to present meaningful insights about the customers perception and restaurants features, in order to improve the rating of the Bangalore Restaurants. As there is a wide number of Restaurants in Bangalore City, there were got several Reviews from Customers and other meaningful information such as the Menu Type, the restaurant type, the location, to improve the customers satisfaction. As it is known, when the businesses or companies work for listening to the customers feedback, the firms can improve as same as their rating. Therefore, the Reviews are highly beneficial to be analysed with a Text Mining Process. Furthermore, there is also a wide range of information gathered from the restaurant that involves a quantitative analysis. Thus, this report analyse the performance of models capable of predicting new data regularly and the one that is more adapted to the business scenario.

As it is known, the rating of restaurants is highly important as it has plenty of benefits not only for the restaurants owners, but it also for the community such as helping to increase the tourism through the visit of the restaurants ,it also attracts more investors who are looking for opportunities as setting up new restaurants new hotels next to the restaurants top rated restaurants.

The main question is: What are the restaurants that the customers most prefer to give the highest rate?

Extension

For instance, as there is a high purpose of improving the Bangalore economy meanwhile the city becomes more attractive to our investors. Even though, it is not a short-term goal to enhance the economy and the metrics of life quality such as urbanization, health, communication, education, security, and others, the popularity of the restaurants is a good starting point for this business problem.

Through improving the performance of the restaurants of the city, the rate of those restaurants would increase.

Thus, the information developed in this Analysis is highly useful for the decisions to be made in the type of restaurants that our investors should invest. As it is explained, there are plenty of benefits after investing in the restaurants recommended as there is more certainty about the return of investment, the revenues and the successfulness of the restaurant. Because the company clients are initially the citizens of Bangalore, the target would be expanded with the successfulness of the restaurants, in which the aim is attracting more tourist that get fascinated with the restaurants.

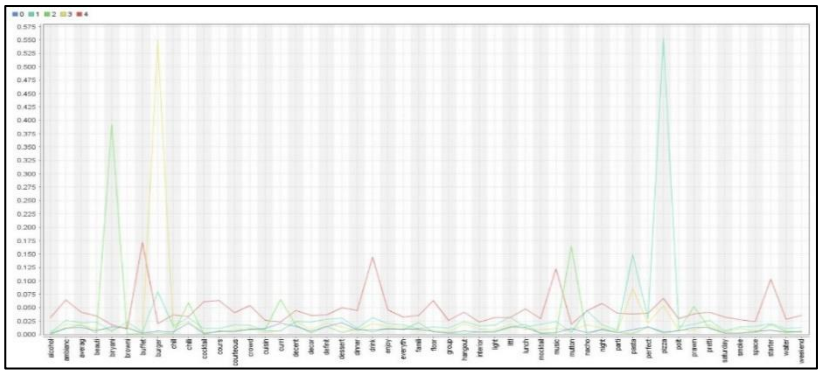
During this analysis, several models were developed to build a rating predictor, taking a sample of 20.000 records out of 40.000. For instance, the models built were Gradient Boosted Tree, a Linear Regression, and Neural Nets, in which there were used custom ensembles to improve the model, cross validation and other Optimization processes that will be explained more in detail during the report.

The model selected is the Gradient Boosted tree, which got the best performance, in which the results obtained were 0.247 for RMSE, 0.166 for MAE and 0.778 in Correlation.

.

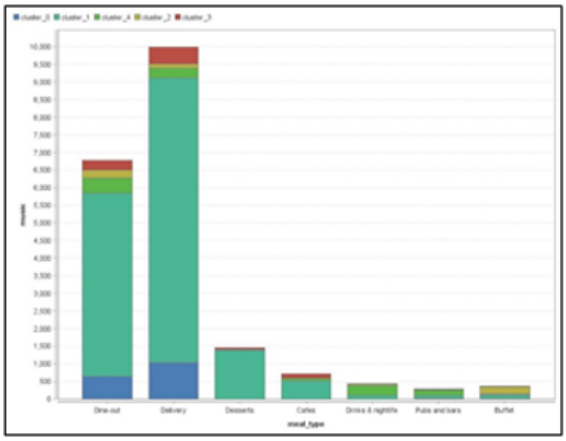
Data exploration and relationships - Clustering in RapidMiner (one page)

Taking into account that a sample data of 10.000 was selected. The process of Parsing involves various sub processes, in which the attributes such as meal type, menu item, restaurant type and reviews text, were chosen to be analysed. Then, due to review, text attribute should be analysed, as it is the attribute that contains only text, was necessary to use the Operator Nominal to text to later on use Weight by Information Gain Operator for developing dimensionality reduction.

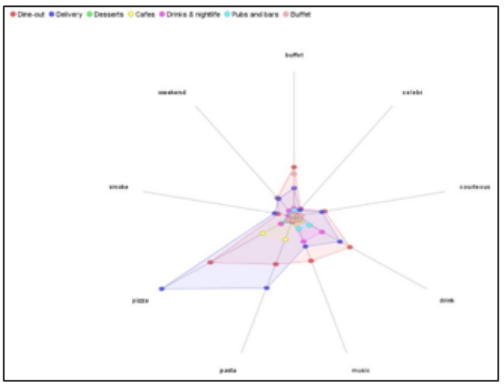


However, there is a main step previous to the dimensionality reduction, which is Process Documents to Data, which has 5 processes inside, since transforming all the review words to lower case to select the words depends on their length, and in between applying tokenization, stemming, and filtering stop words. Then, it is necessary to select the weights to apply the cluster operator for segmentation with 5 k, specified from cluster 0 to 5. The following plot in the Figure 2.2. represents the five Clusters generated after the text mining and clustering:

For instance, lets follow the word “music”, selected as a reference, as it can be seen the cluster 4 identified with red colour, represents the best cluster for this word with 0.123. After getting this results with a cluster of 5 K, a cluster of 10 K and 20 k were also tested, in which the word “music” has the highest representation in the cluster 2 with 0.217, and with 20 K, music has the highest representation with the cluster 7 with 0.284. It means that the strength of the word in the clusters change, when the parameters change.

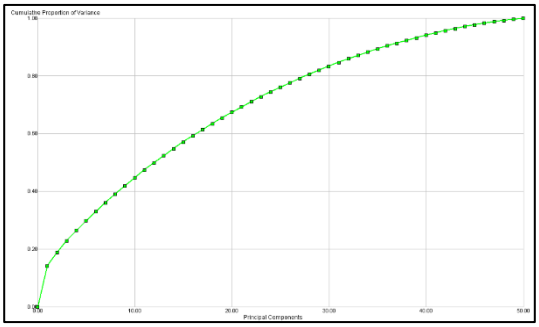


The figure 2.3 shows the representation of the word “music” in the meal types attribute. Thus, as it can be seen, the word music has the highest value in Delivery with the Cluster 1, and the lowest representation in Buffet, according to this, most of the customers use music in Delivery meal type in their reviews about the restaurants.



This web chart, figure 2.4, represents the relationship between the different meal types and the words used for the customers in their reviews, as it can be seen, the word Pizza is the most used when they are talking delivery, dine-out and cafes. As this chart is suitable for showing multivariate data in a chart with two dimensions, the word music is also represented to give a better understanding of the frequency of the word in the meal types attribute.

For instance, for answering the question A, the groups of similar restaurants are the 3 lowest represented in the graph, which have same values in the counting of the 5 clusters applied, those are the restaurants with the following meal types: 1. Drinks and night life. 2. Pubs and Bars. 3. Buffet



Extension:
A cluster of 10 k and a cluster of 20 k were also applied for testing the data. Thus, this diagnostic chart was represented through performing PCA after developing the Knn Global Scoring. The graphs denotes the main two component Analysis data points as it can be identified following the Axis X of PCA and the Axis Y of Cumulative Proportion of Variance.

Data exploration and cleanup - Anomalies in RapidMiner (one page)

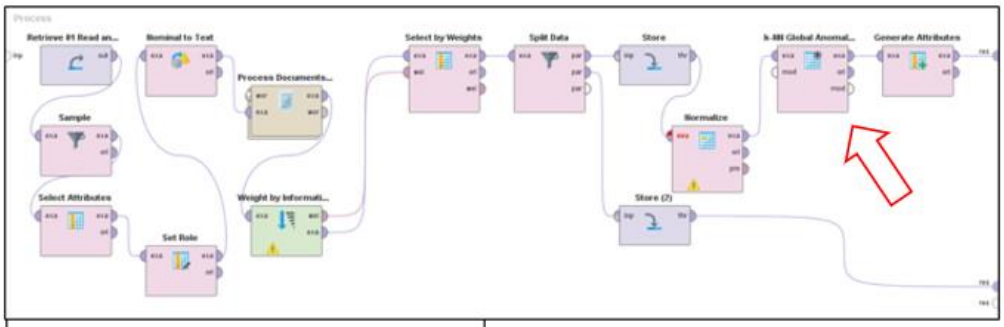


Figure 2.1: Process of Anomaly detection with Knn

Expectation

During this process, was necessary two split in two parts, training and testing, the data collected after the text mining and segmentation process. The training dataset is normalised and the process of knn is used for representing the clusters or neighbours, showing the data with similarities for a better understanding.

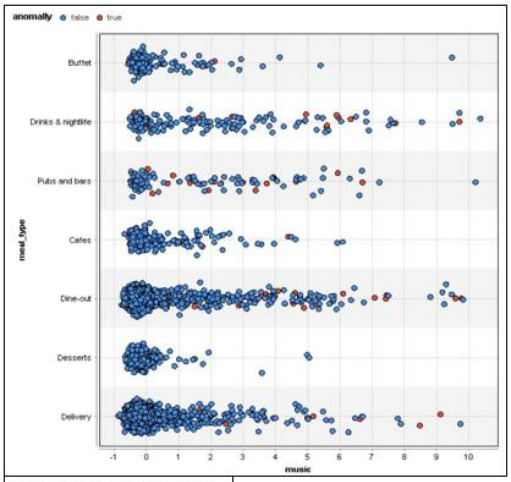


Figure 2.2: Anomaly detection

At the end, to generate the attributes, the outliers or abnormal data is filtered to represent the data with the highest anomalies, in order to understand better the visualization. Thus, this scatter plot shows the clustering visualization, in which the blue data points are the data next to the average, or representative data, and the red data points are the anomalies. For this visualization, I took the word “music”, from the Customers Review, mix of textual and structured data, to represent the frequency of this word in the seven meal types.

As it can be identified, in dine-out the word was used more times than in the others meal types, on the other hand, desserts meal type has the lowest use of that word. For instance, it can be identified that the restaurants categorised as dine-out have played more music than the other restaurants of the city.

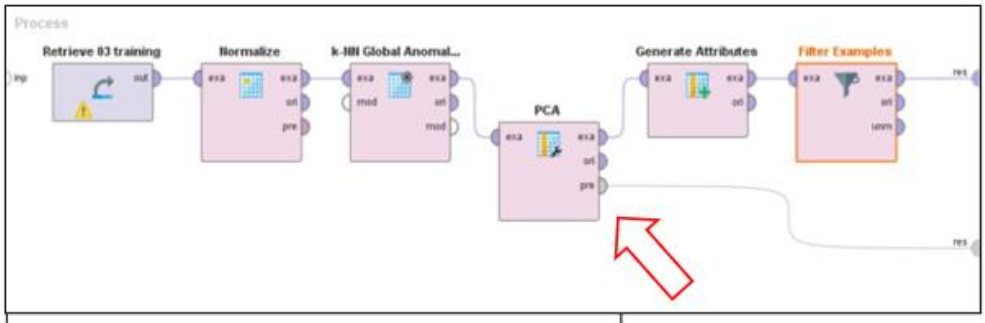


Figure 2.3: Process of Anomalies removed and visualised with PCA

Extension

In the figure 2.3, it is demonstrated that for developing the PCA process, it has to be used after the knn process, which is also applied for visualising the two principal components as these have the largest variance. Finally, the outliers were filtered and represented in the figure 2.4.

For instance, after understanding the anomalies in the dataset, it is necessary to use PCA or Principal Component Analysis to identify the variance, or the distance of some data points from the most frequent data points. In this scatter plot, the two first Principal Components are represented, and as it can be seen, the outliers have been removed as those affect significantly the mean or the average and in the standard deviation or the spread of the data from the average. Thus, if we want to analyse an attribute and there are outliers, those impact in a misunderstanding of the information, in this case the reviews, and might affect the prediction model significantly, giving wrong insides about the feedback of the customers, affecting the decisions that have to be made. Thus, the anomalies were identified and removed.

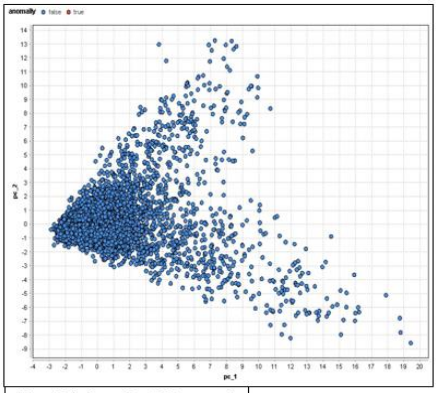


Figure 2.4: Anomalies PCA removed

Create a Model(s) in RapidMiner (two pages / page 1)

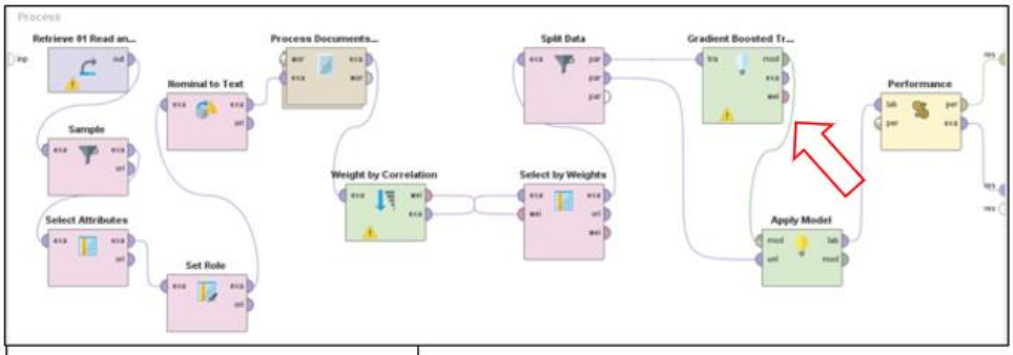


Figure 3.1: Gradient Boosted Three, GBTs

Expectation

The second Model implemented was Gradient Boosted Tree, which is a supervised learning algorithm based in the decision tree model. The parameters used involved the learning rate as 0.1, the number of trees that for default are 100 the maximal depth that for default is 10. The data is divided in two parts, for training with 0.7 and testing with 0.3.In the model there are tree metrics that identify the accuracy of the continuous data analysed, the Root Mean Squared Error (RMSE), the Absolute Error (MAE) and the Correlation.

Over the tree metrics measured, although RMSE gives a higher weight to large errors and there were identified several anomalies or outliers, MAE gives the average of the difference of observed and predicted values and do not penalise the higher differences or predictions. Thus, MAE is more suitable than RMSE.

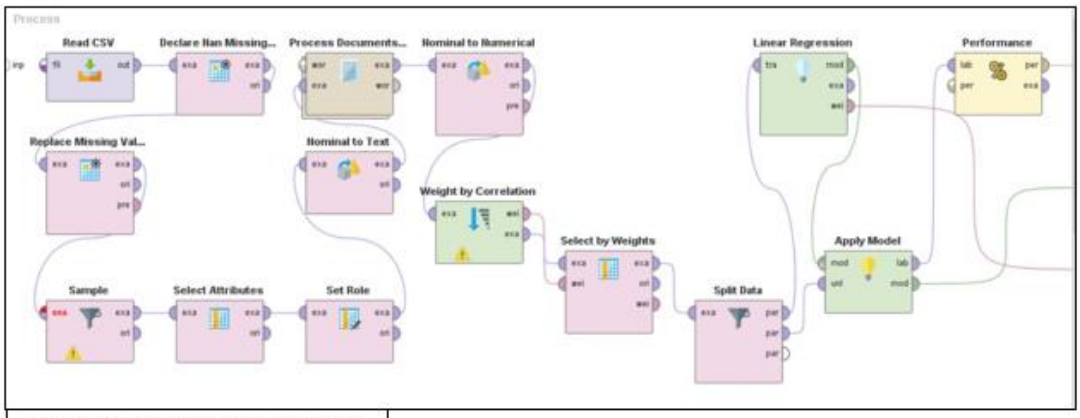


Figure 3.2: Linear Regression Model

For creating the Linear Regression model, as it can be seen in the figure 3.2, the missing values from the sample data of 10.000 records, were declared and replaced for the average, mean for categorical data and mode for numerical values, then the attributes selected and the text data is transformed as was previously indicated.

The data had to be transformed through dummy variables, using Nominal to Numerical Operator, as this model only works for numerical data, and then the data is split in 70% and 30%, as was previously mentioned. The parameters indicated in Linear Regression involved the min tolerance that is 0.05 and the feature selection as M5 Prime.

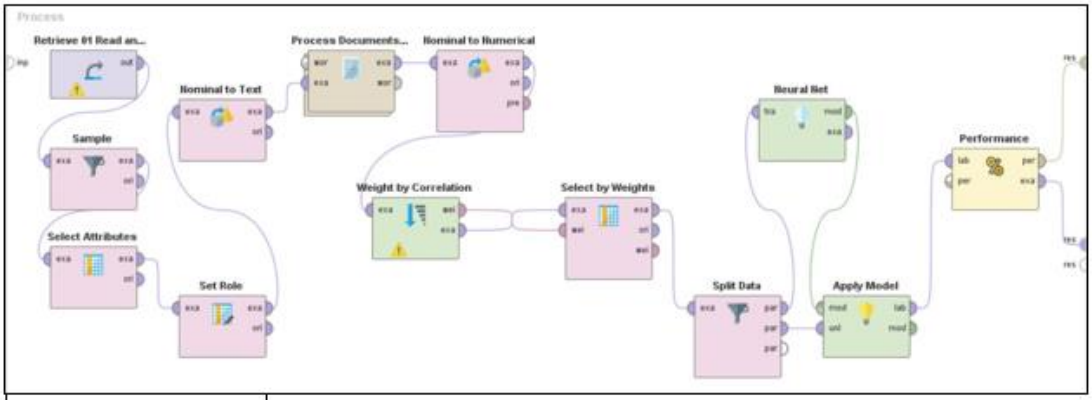


Figure 3.3: Neural Nets

For the Third Model, which is Neural Nets, the data was also divided in 70% and 30%. The parameters indicated were 200 in training cycles, .0.01 in learning rate, that determines in every step how much the weights are changed. Furthermore, although there was used a Gaming Laptop brand Asus GL, with 16 G.B the Ram and with a Graphic Card Nvidia GTX960 and with 2.60HZ of processor, a sample of 10.000 records had to be used for running the processes. For instance, as Neural Nets runs several training cycles and a wide amount of processing units using Numerical values transformed through dummy coding, the processes consumes resources such as time.

Create a Model(s) in RapidMiner (two pages / page 2)

Thus, the best model performed was Gradient Boosted Three with a MAE of 0.197, a RMSE 0.275 of and a Correlation of 0.727.

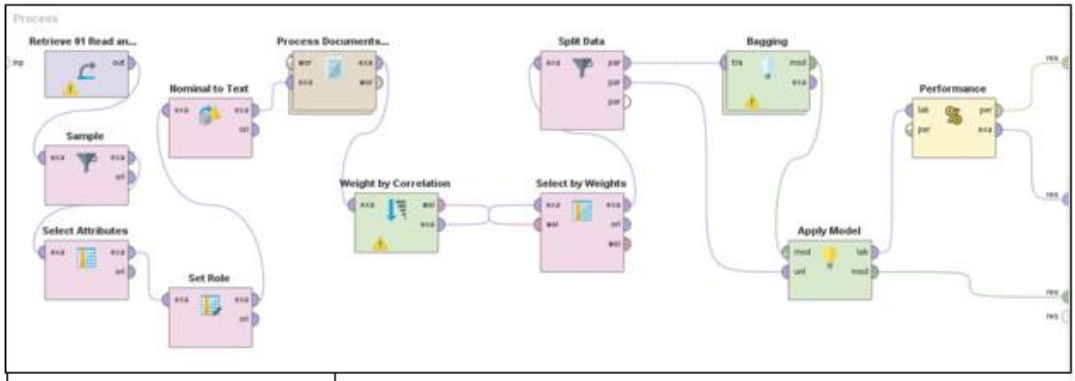


Figure 3.4: Custom Ensembles

Extension

The four model implemented was Custom Ensemble that followed the same processes of the previous models until the data is split in 70% and 30%, then there was used the Bagging Operator or Bootstrap Aggregation, which is one of the Ensembles.

Bagging validates several models with different parameters that previously took different random samples and gives an average performance as the output, in order to reduce the variance in prediction. The parameters defined in Bagging were a simple ratio of 0.9, which defines the fraction of examples used, with 10 iterations. As the custom ensembles aims to reduce the variance, a model such as Random Forest or Gradient Boosted Tree has to be inside the Operator. Thus, the GBTs with the previously default parameters mentioned, was used inside Bagging.

Evaluate and Improve the Model(s) in RapidMiner (two pages / page 1)

Expectation

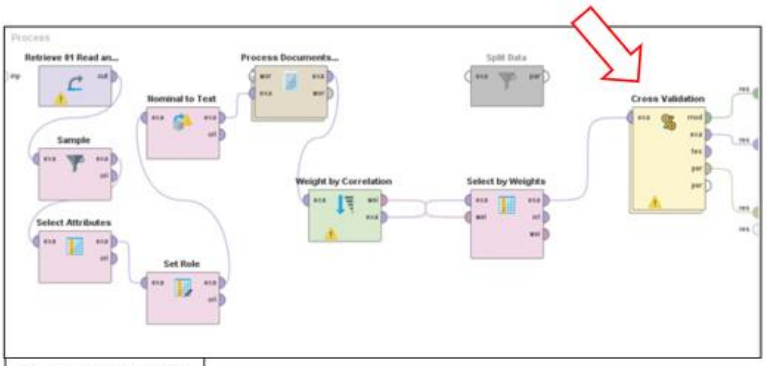


Figure 4: Cross Validation

The following figure 4 represents the process run in all the models for optimization.

GBTS									
K- Folds	5			10			20		
Number of trees	20	50	80	20	50	80	20	50	80
RMSE	0.366	0.333	0.312	0.366	0.311	0.333	0.366	0.333	0.311
MAE	0.268	0.248	0.234	0.267	0.233	0.248	0.267	0.248	0.233
Correlation	0.670	0.692	0.703	0.675	0.706	0.694	0.674	0.695	0.707

Figure 4.1: Cross Validation of Gradient Boosted Tree

Due to the fact that Cross Validation was used to optimize the model, it was tested with 3 different number of folds, 5, 10 and 20 folds. Moreover, the model was tested with 3 different number of trees such as 20,50 and 80 in every number of folds, in order to identify the best model. Thus, the best GBTs was the one with the 20 number of folds and 80 number of trees, with a RMSE of 0.311, a MAE of 0.233 and a Correlation of 0.707, which also was the model with the best performance between the others 4 models. Refer to figure 4.1.

Linear Regression			
K- Folds	5	10	20
RMSE	0.317	0.316	0.316
MAE	0.238	0.238	0.238
Correlation	0.597	0.597	0.597

Figure 4.2: Cross Validation of Linear Regression

For Linear Regression, after performing Cross Validation, the results obtained were highly similar in all the test developed, using folds of 5, 10 and 20, were the best model had 10 and 20 folds. Thus, the best RMSE was 0.316, MAE of 0.238 and Correlation of 0.597. Refer to figure 4.2.

Neural Nets									
K- Folds	5			10			20		
Training Cycle	200	600	800	200	600	800	200	600	800
RMSE	0.318	0.314	0.311	0.331	0.326	0.324	0.321	0.317	0.316
MAE	0.244	0.240	0.238	0.251	0.247	0.245	0.240	0.237	0.236
Correlation	0.624	0.636	0.636	0.625	0.636	0.639	0.619	0.636	0.638

Figure 4.3: Cross Validation of Neural Nets

After running the Cross Validation for Neural Nets, and tested 9 times with 5, 10 and 20 folds and inside every fold with 200, 600 and 800 training Cycles, the best performance model was the one with 10 folds and 800 training cycles. This model got a RMSE of 0.324, MAE of 0.245 and a Correlation of 0.639, as it is demonstrated in the figure 4.3.

After applying Cross Validation to Custom Ensemble, it got a RMSE of 0.312, a MAE of 0.232 and a Correlation of 0.701, as it can be seen in the figure 4.4.

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.312 +/- 0.013 (micro average: 0.312 +/- 0.000)
absolute_error: 0.232 +/- 0.010 (micro average: 0.232 +/- 0.209)
correlation: 0.701 +/- 0.031 (micro average: 0.701)
```

Figure 4.4: Cross Validation of Custom Ensemble

Evaluate and Improve the Model(s) in RapidMiner (two pages / page 2)

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----root_mean_squared_error: 0.299 +/- 0.000
-----absolute_error: 0.228 +/- 0.193
-----correlation: 0.631
]
Neural Net.training_cycles       = 1150
Neural Net.learning_rate         = 0.010000000000000002
Neural Net.momentum      = 0.05
```

Figure 4.5: Optimization Grid of Cross Validation of Neural Nets

Extension

After using Optimization Grid for the Cross Validation of Neural Networks, the results obtained were 0.299 for RMSE, 0.228 for MAE and 0.631 in Correlation. As it is known, Optimization Grid takes into account the subproccesses of all the combinations of the parameters indicated. Thus, as it can be identified in figure 4.5, the best parameters for training cycle, learning rate and momentum are indicated.

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----root_mean_squared_error: 0.247 +/- 0.014 (micro average: 0.248 +/- 0.000)
-----absolute_error: 0.166 +/- 0.009 (micro average: 0.166 +/- 0.183)
-----correlation: 0.778 +/- 0.023 (micro average: 0.778)
]
Gradient Boosted Trees.number_of_trees  = 63
Gradient Boosted Trees.maximal_depth    = 100
Gradient Boosted Trees.learning_rate    = 0.1
```

Figure 4.6: Optimization Grid of Cross Validation of Gradient Boosted Tree

Furthermore, after using Optimization Grid for the Cross Validation of the Gradient Boosted Tree, the results obtained were 0.247 for RMSE, 0.166 for MAE and 0.778 in Correlation. Thus, for this process of Grid Optimization a sample data of 10.000 records was used.

For instance, the results obtained in the Gradient Boosted tree, are the best results obtained. As it can be seen in the figures 4.8 and 4.9

Optimize Parameters (Grid) (245 rows, 5 columns)

iteration	Gradient Boosted Trees.number_of_trees	Gradient Boosted Trees.maximal_depth	Gradient Boosted Trees.learning_rate	root_mean_squared_error
1	10	1	0.100	0.368
2	28	1	0.100	0.348
46	10	34	0.233	0.337
16	10	51	0.100	0.344
31	10	100	0.100	0.344
3	45	1	0.100	0.331
4	63	1	0.100	0.326
62	28	84	0.233	0.338
47	28	34	0.233	0.338
17	28	51	0.100	0.333
123	45	51	0.500	0.369
5	80	1	0.100	0.324
32	28	100	0.100	0.333
6	10	18	0.100	0.345
54	63	51	0.233	0.351
48	45	34	0.233	0.334
7	28	18	0.100	0.330
18	45	51	0.100	0.328

Figure 4.8: Matrix of Iterations in Optimization Grid of GBTs

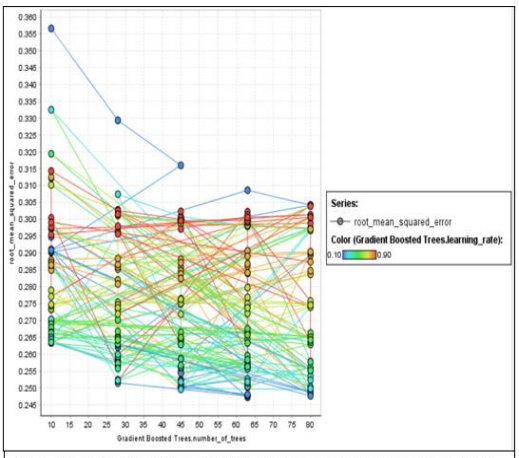


Figure 4.9: Advanced Chart Model GBTs with Cross Validation and Optimization Grid

This figure 4.9 indicates the number of trees in the Axis X and the RMSE in the Axis Y, the datapoints are coloured with the Learning rate. The RMSE is higher when the number of trees is only ten, however the learning rate is higher when the number of trees is small

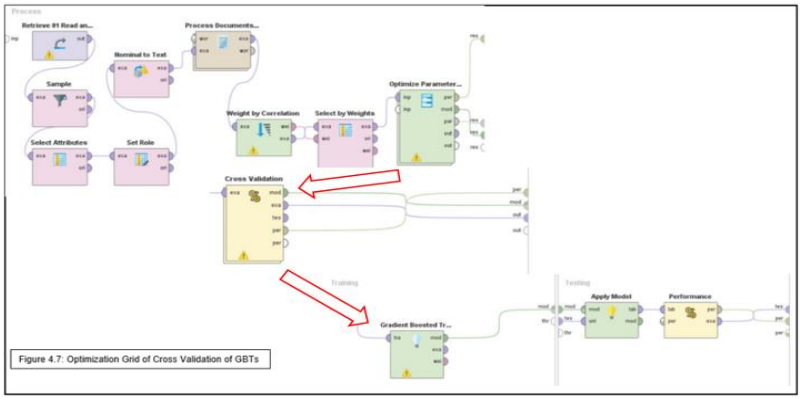


Figure 4.7: Optimization Grid of Cross Validation of GBTs

The following figure 4.7 shows the processes inside of Optimization Grid, which can be followed with the red arrows. In the Optimization Grid, fail on error was selected in the parameter of error handling and the enable parallel execution. Cross Validation, which is inside the Optimisation grid, remains with the 20 number of folds and inside Cross Validation

Provide an Integrated Solution in RapidMiner (one page)

Expectation

Over the models developed, Linear Regression, Gradient Boosted Tree, Neural Nets and Custom Ensemble, which were optimized using Cross Validation and Grid Optimization for optimizing the hyper- parameters, **the best model is the Gradient Boosted Tree or GBTS**, as it was already defined. In this model, the trees took from one decision tree, are developed in order to minimise the error and correct the predictions, moreover, this model is suitable for this sort of datasets with a countable number of attributes rather than Neural Nets that mainly performs better with a wide number of attributes in a dataset.

For instance, **the best strategy for predicting the rating is with the implementation of GBTs Model with the Hyper parameters already set**. To demonstrate the performance of this model, in this section the new data called deploy data is applied to the Model and the output will be demonstrate the effectiveness of this model.

Extension

Pre-processing models, best predictive model, as well as, word lists, weights and PCA models saved for deployment. Saved models retrieved and applied to new data. Shown that a single enquiry handled.

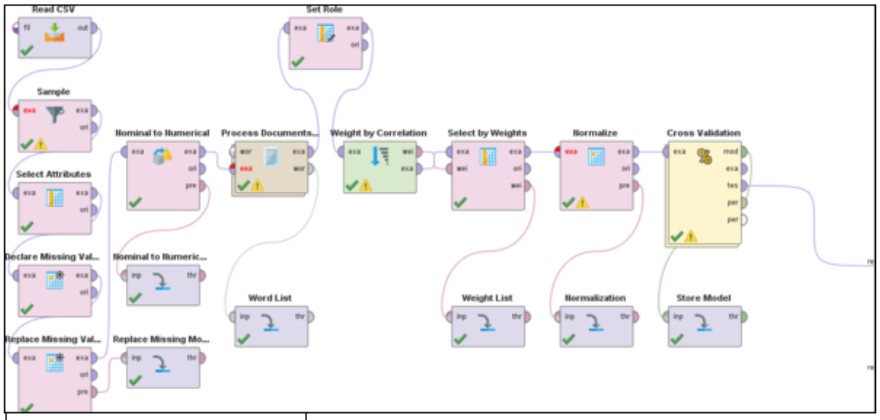


Figure 6.1: Model Deployment

The next steps need to be followed in order to open and load the CSV File:
a) Download the Zip File, unzip it in your PC Disk, preferably in Disk (C:).
b) Create a Repository Folder in RapidMiner, in the case that there was not a Repository already Created.
c) Type in the Operator Tab "Read CSV" to read the File created for the model. Moreover, to see the process, open the process called "Model Deployment Final" located in Repository tab.

The process developed during this project are stored as: Read Tomato Feedback, Cluster 1 Rest Type, Cluster 2 Meal Type, all the process that start with the name Anomaly, mainly Anomaly Final and Anomaly Detection PCA Plot, all the processes that start with M such as M1, M2, M3 and M4 as those were the models created and optimized with Cross Validation and Grid Optimization, and the model called Further Research that gives another way to optimize the model.

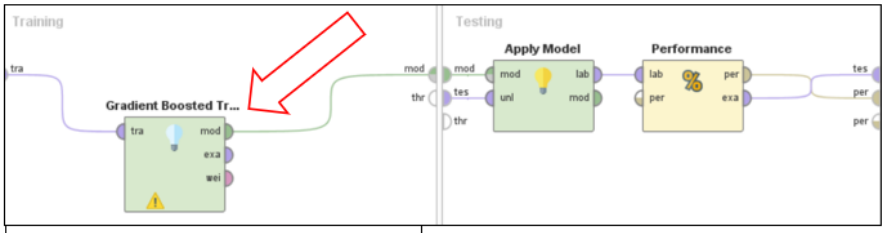


Figure 6.2: Cross Validation in Model Deployment

As it can be seen in the figure 6.2., it follows the same process that was done for all the models with a sample of 10.000 records, about pre-processing, text mining, weighting attributes, normalising and ending up with the Cross Validation process. Inside the Cross Validation, the best model with the Optimum Hyper Parameters were implemented, in order to use it in the new data.

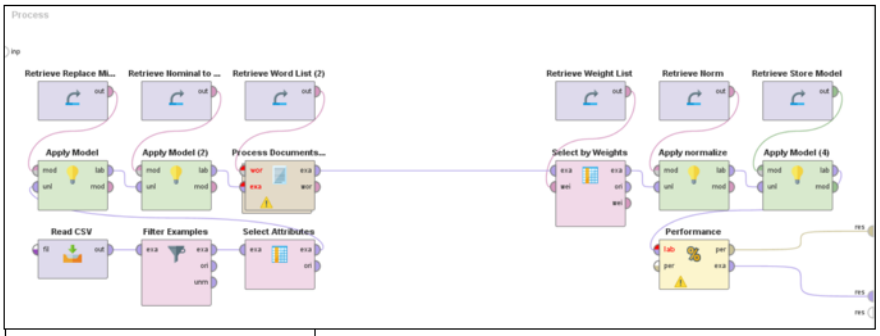


Figure 6.3: Deployment in New Data

Thus, in the figure 6.3, **the model is applied in the New Data** after retrieving all the information stored, which adds the attributes in the deploy dataset, to use in the prediction according to the business case.

Further Research and Extensions in RM (one page)

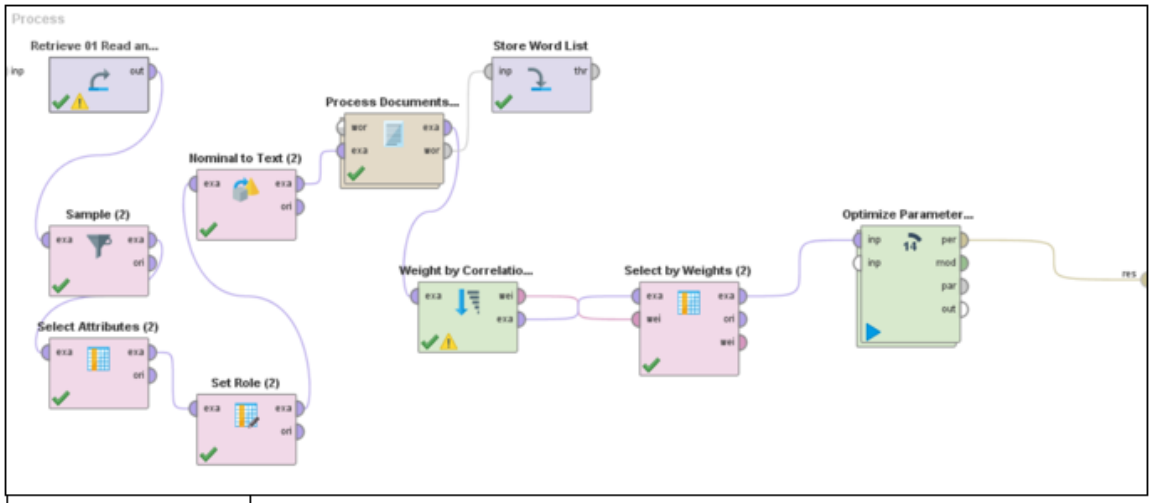


Figure 6.1: Stacking

Expectation

The figure 6.1 shows the model developed for optimizing the performance using Stacking of Custom Ensemble, when Bagging was already used, now the model is tested with stacking.

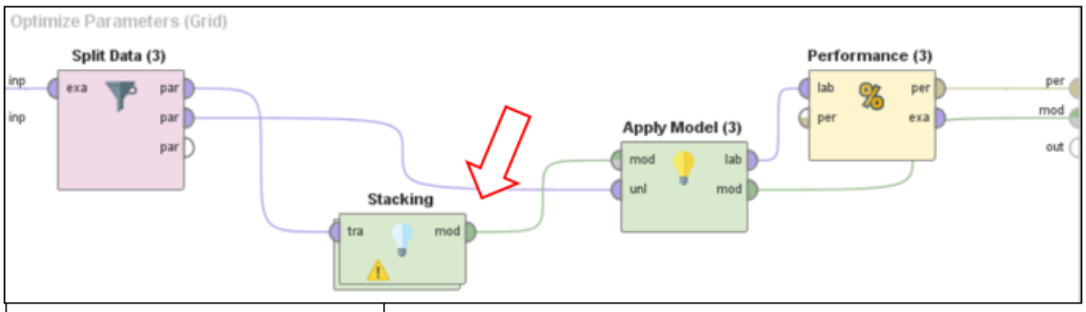


Figure 6.2: Optimization of Stacking

Therefore, the model follows the same processes in the pre-processing than the others model implemented, using Grid Optimization. As it can be seen in figure 6.2, inside Grid Optimization, the data is split and then the stacking Operator is applied, and the testing part is applied to the model performance directly to test with the data trained.

Figure 6.3: Inside Stacking

In the figure 6.3, it is used GBTs and K-nn to inside stacking and then put into the model to check the performance.

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.312 +/- 0.001 (micro average: 0.312 +/- 0.000)
absolute_error: 0.234 +/- 0.002 (micro average: 0.234 +/- 0.207)
correlation: 0.703 +/- 0.009 (micro average: 0.703)
```

Extension

As it can be seen and as it is certain, the Gradient Boosted Tree was optimised using Stacking. Thus, the RMSE is 0.312, the MAE is 0.234, and the Correlation is 0.703. Nevertheless, the GBTS with Grid Optimization got a higher correlation and less error in RMSE.