Assignment A2: Text Mining + DT + Neural Nets + Optimisation			
Student Name *(as per record)*	**Shantanu Gupta**	**Student No**	*218200234*
My other group members		**A2 Group No**	*As per Cloud Deakin group number*
Student Name *(as per record)*		**Student Nos**	*Student number*
			Student number
			Student number

	Exceptional	Meets expectations	Issues noted	Improve	Unacceptable
Exec Report					
Create Models					
Evaluate & Improve					
Provide Solution					
Research & Extend					

Brief Comments

Read these notes as we are really trying to help you out!
Remember: If it is not in this report, it does not exist and does not get marked!

You can use the above form to estimate the expected mark against the rubric (see the assignment "info" document). Be realistic and note that we will find many problems you may not be aware of.

Assume that markers may be tired when assessing your work and they may miss some important aspects of your submission when not presented clearly, or when you deviate from the structure of this template, or if you do not include them in your report. So be clear, number all tables, charts and screen shots used as evidence, describe all visuals, cross-reference your analysis with evidence.

Submit this report in **PDF format** to avoid accidental reformatting of the content.
Submit all RapidMiner processes (.RMP files) in a separate **ZIP archive**, so that if there is any doubt we could load your work and replicate your results (we will not do this to find missing report parts).

Ensure that the report is readable and the font is no smaller than Arial 10 points. In the report include only the most significant results for your analysis and recommendations.

You will be able to submit your work once only so make sure you get it right – check these before posting on CloudDeakin: Is this your document? Is this the correct unit, assignment, year and trimester? Is your name entered above? Is the group number included and is it correct? Are names of your group members entered as well? Are all pages included? Does it all fit into the required page limit? Have you zipped all RapidMiner files (.RMP files)? Is the report contents yours alone?

Then after the submission – check these: Has the PDF report been submitted? Has the Zip archive of RMP files been submitted? Can you retrieve and reopen both back from your submission folder?

Note that the late penalty will be calculated on the date and time of the last submitted file.

Finally, as all reports will be inspected for plagiarism, ensure that your analysis, your evidence, your way of thinking, your report and its presentation are unique and demonstrate your ability to create it all independently. So if you work in a team compare your submission to those of your team members and make it quite distinct in both contents and form. Any part of this report that bears any resemblance to another students' report or any information source written by others or by you for another unit (e.g. on the web) will be treated as plagiarism.

Total

Include: Report and RMP files, with clear comments supplied to (easily) reproduce reported results.

Executive summary (one page)

Expectation

Australian Wine Importers (AWI) give data that contains 130k of wine reviews, where each description review is written by a sommelier, and give an 80–100 point rating. Wine Enthusiast ranks wines on a 100 point scale with only 80+ point wines receiving a written review. According to one website post, the scores roughly correspond to:

- Classic 98–100: The pinnacle of quality.
- Superb 94–97: A great achievement.
- Excellent 90–93: Highly recommended.
- Very Good 87–89: Often good value; well recommended.
- Good 83–86: Suitable for everyday consumption; often good value.
- Acceptable 80–82: Can be employed in casual, less-critical circumstances

Since each wine should have been expertly described by someone trained in the art of wine tasting, AWI thought we could use the data to analyse wine description. In Wine Reviews, the wine description plays a vital role. The "description" of each of these reviews is comprised of a few sentences in which the wine reviewer describes what he/she tastes in the wine and gives an opinion about the wine. A good description can make your wine stand out. It also helps get reviews faster. The objective of this project is to build a machine learning so that we can build a predictive model to estimate (rating) of a wine belonging based on descriptive words by sommelier? Lastly, it will help you get a prediction of wine points and also, sound smarter when I review my next one. Let's see what we can find in the wine description. We will use Rapid Miner for handling this project.

Extension

Below is a brief introduction of the steps that were undertaken by using Rapid Miner to analyse our data:

- Identify the problem
- Data Preparation and analysis
- Data Exploration and pre-processing
- Applying different machine learning algorithms (Effort to improve parameters)
- Prediction Model development and deployment
- Prediction Model Implementation

The end goal for this project is to understand the input processes of importing wine so as to ensure that we get not just the good quality of the wine, but it also should be price effective, highest points and also reduces some kind of dependencies over expensive sommeliers. Any effort to improve this situation must consider the revolution in these industries and should leverage existing data and information as extensively as possible. However, with a robust implementation, AWI will have the right machine intelligence in place to the gut instinct of sommelier and became competitive in the market.

All my models have these selected parameters:

Unstructured Column (Text, A1)-description, Title

Structured Column (A2) - Variety, winery, country, province, designation

A mixture of Structured and Unstructured Column (A3) - Country, description, designation, province, title, variety, winery

Also in my analysis, I have not changed random seed, I used computer default one.

References:-

The official 2019 Wine Vintage Chart (2019), https://bit.ly/2F7CUDH
Exploring and Classifying Wine Enthusiast Reviews, John, 2017, https://bit.ly/2VBNfk3
Introduction to Natural Language Processing (NLP), Morgan, 2018, https://bit.ly/2UNfTtZ
Predicting Wine Quality using Text Reviews, Feb, https://bit.ly/2ZJrjTt
Medium (2018), Cancer Dabakoglu, Wine reviews Visualization & NLP, https://bit.ly/2Uq2llP
Medium (2018), Wine Reviews, https://bit.ly/2WW7zKv
Sisense (2018), Elana Roth, 5 Steps to Data-Driven Business Decisions, https://bit.ly/2hjoB1g
Datacamp.com (2018), Duong Vu, Generating word clouds in python, https://bit.ly/2EosS03

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

Expectation

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.455 +/- 0.000
absolute_error: 1.945 +/- 1.498
relative_error: 2.21% +/- 1.73%
squared_error: 6.028 +/- 8.769
correlation: 0.584
squared_correlation: 0.341
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.174 +/- 0.000
absolute_error: 1.702 +/- 1.352
relative_error: 1.93% +/- 1.55%
squared_error: 4.725 +/- 7.157
correlation: 0.712
squared_correlation: 0.507
```

Figure: Performance in Decision Tree and Gradient Boosted Trees for A1

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.456 +/- 0.000
absolute_error: 1.961 +/- 1.479
relative_error: 2.22% +/- 1.71%
squared_error: 6.033 +/- 8.748
correlation: 0.600
squared_correlation: 0.360
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.407 +/- 0.000
absolute_error: 1.920 +/- 1.452
relative_error: 2.18% +/- 1.67%
squared_error: 5.793 +/- 8.359
correlation: 0.614
squared_correlation: 0.377
```

Figure: Performance in Decision Tree and Gradient Boosted Trees for A2

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.432 +/- 0.000
absolute_error: 1.944 +/- 1.461
relative_error: 2.21% +/- 1.68%
squared_error: 5.915 +/- 8.540
correlation: 0.598
squared_correlation: 0.358
```

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.178 +/- 0.000
absolute_error: 1.704 +/- 1.357
relative_error: 1.93% +/- 1.56%
squared_error: 4.743 +/- 7.248
correlation: 0.711
squared_correlation: 0.505
```

Figure: Performance in Decision Tree and Gradient Boosted Trees for A3

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.795 +/- 0.323 (micro average: 2.813 +/- 0.000)
absolute_error: 2.262 +/- 0.313 (micro average: 2.262 +/- 1.673)
relative_error: 2.55% +/- 0.33% (micro average: 2.55% +/- 1.88%)
squared_error: 7.915 +/- 1.878 (micro average: 7.915 +/- 10.772)
correlation: 0.588 +/- 0.018 (micro average: 0.477)
squared_correlation: 0.346 +/- 0.021 (micro average: 0.227)
```

PerformanceVector

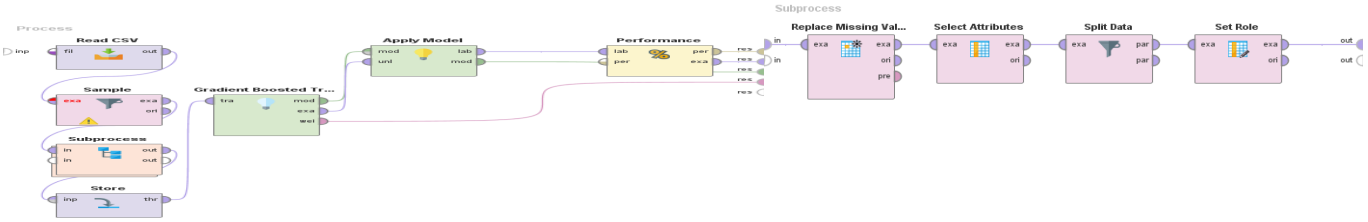
```
PerformanceVector:
root_mean_squared_error: 2.891 +/- 0.493 (micro average: 2.933 +/- 0.000)
absolute_error: 2.331 +/- 0.424 (micro average: 2.331 +/- 1.780)
relative_error: 2.66% +/- 0.51% (micro average: 2.66% +/- 2.08%)
squared_error: 8.601 +/- 3.100 (micro average: 8.601 +/- 12.570)
correlation: 0.592 +/- 0.015 (micro average: 0.451)
squared_correlation: 0.351 +/- 0.018 (micro average: 0.203)
```

```
PerformanceVector:
root_mean_squared_error: 2.876 +/- 0.464 (micro average: 2.914 +/- 0.000)
absolute_error: 2.323 +/- 0.439 (micro average: 2.323 +/- 1.759)
relative_error: 2.62% +/- 0.47% (micro average: 2.62% +/- 1.98%)
squared_error: 8.488 +/- 2.956 (micro average: 8.491 +/- 12.035)
correlation: 0.564 +/- 0.047 (micro average: 0.463)
squared_correlation: 0.320 +/- 0.053 (micro average: 0.214)
```

Figure: Performance of Neural Network forA1, A2, and A3

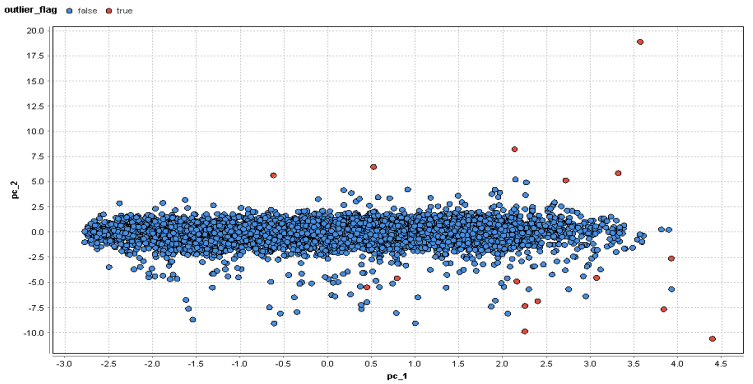
Create a Model(s) in Rapid Miner (two pages / page 2)

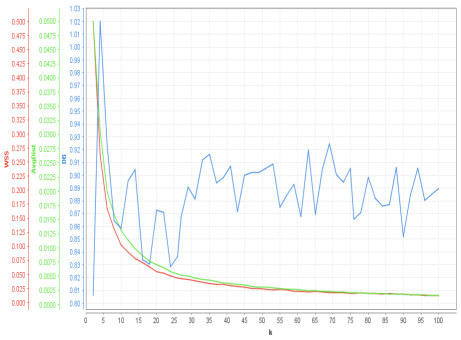
- The root-mean-square error (RMSE) is a measure of how well my model performed. For all my A1, A2, A3 models GBT outperformed in RMSE as compared to the decision tree. A similar trend is followed by Mean absolute error.
- Neural Network performance is not as good as I trained the data for only 15000 samples. As data is not enough that's why performance is not good. All my neural networks are worse in performance as compared to the decision tree.
- Most of the correlation value is greater than 0.5 that shows there is a positive, moderate and moderate relationship between predictor and response variable (points).



- I have used 15000 samples. In the subprocess, I used to replace missing values in which I replace all missing values by mean (numerical), mode (categorical). Selection of attributes is done on the basis of the structured and unstructured column. In any of my model I have not used id, region_1, region_2, taster name, and taster twitter handle because id, taster name, taster twitter handle means no practical significance in predicting points, and region because there are so many missing values. In the selection of attributes, I already mention in exec report of my different models. Besides that, I have also used points and price. I have used set role in which I used points as a label. The predicted results are mentioned above.

Extension





Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.495	0.248	0.248
PC 2	1.290	0.185	0.433
PC 3	1.187	0.157	0.590
PC 4	0.995	0.110	0.700
PC 5	0.960	0.102	0.802
PC 6	0.887	0.087	0.890
PC 7	0.731	0.059	0.949
PC 8	0.669	0.050	0.999
PC 9	0.101	0.001	1.000

ExampleSet (15000 examples, 1 special attribute, 9 regular attributes)

Row No.	score	country	descrip...	designa...	points	price	province	title	variety	winery
10363	6.252	France	This new...	Le Mesni...	100	617	Champa...	Salon 20...	Chardon...	Salon
12940	6.252	France	Almost b...	Reserve	100	1500	Bordeaux	Château ...	Bordeau...	Château ...
13291	5.951	Portugal	A powerf...	Nacional...	100	650	Port	Quinta d...	Port	Quinta d...
6480	5.951	Portugal	This is th...	Barca Ve...	99	426	Douro	Casa Fe...	Portugue...	Casa Fe...
4172	5.951	France	Pure Ch...	Clos du ...	99	800	Champa...	Krug 200...	Chardon...	Krug
4500	5.650	Italy	Here's a ...	Messorio	99	320	Tuscany	Le Macc...	Merlot	Le Macc...
12648	5.519	France	With its g...	Cuvée Sl...	98	305	Champa...	Pol Rog...	Champa...	Pol Roger
12941	5.474	France	A hugely ...	Reserve	100	359	Bordeaux	Château ...	Bordeau...	Château ...
13184	5.419	France	This isn't...	Reserve	97	800	Bordeaux	Château ...	Bordeau...	Château ...
5636	5.394	France	A beautif...	Reserve	98	319	Burgundy	Clos de ...	Pinot Noir	Clos de ...
4816	4.973	France	Even mo...	Cuvée C...	99	159	Loire Val...	Domain...	Chenin ...	Domain...
3105	4.942	France	This is a ...	Reserve	97	184	Burgundy	Louis Ja...	Chardon...	Louis Ja...
7506	4.917	France	This mo...	Reserve	96	496	Burgundy	Roche d...	Pinot Noir	Roche d...
13185	4.917	Italy	Monte Lo...	Vigneto ...	96	470	Veneto	Dal Forn...	Corvina, ...	Dal Forn...

- I have chosen the outlier threshold>4. I can see from the graph there are 20 outliers. Most of the outliers are in the fringes.
- First two principal components contain 44% of data. We need 5 principal components to capture 80% of the data.
- Using the elbow method (i.e. Within Sum of squares), an ideal number of clusters should be 20.
- Histogram-based scoring represents outlier score, green and red colour. Green means good and red means bad. We can check the individual observation to see unusual behaviour in the data. Visual representation of outlier.

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

Expectation

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.367 +/- 0.047 (micro average: 2.368 +/- 0.000)
absolute_error: 1.888 +/- 0.038 (micro average: 1.888 +/- 1.429)
relative_error: 2.14% +/- 0.04% (micro average: 2.14% +/- 1.65%)
squared_error: 5.606 +/- 0.225 (micro average: 5.606 +/- 8.151)
correlation: 0.627 +/- 0.015 (micro average: 0.626)
squared_correlation: 0.393 +/- 0.019 (micro average: 0.392)
```

Figure: 5-fold cross validation of GBT of model A1

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.447 +/- 0.019 (micro average: 2.447 +/- 0.000)
absolute_error: 1.949 +/- 0.016 (micro average: 1.949 +/- 1.481)
relative_error: 2.21% +/- 0.02% (micro average: 2.21% +/- 1.71%)
squared_error: 5.990 +/- 0.092 (micro average: 5.990 +/- 8.654)
correlation: 0.611 +/- 0.014 (micro average: 0.611)
squared_correlation: 0.373 +/- 0.017 (micro average: 0.373)
```

Figure: 5-fold cross validation of GBT of model A2

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.388 +/- 0.028 (micro average: 2.388 +/- 0.000)
absolute_error: 1.910 +/- 0.032 (micro average: 1.910 +/- 1.432)
relative_error: 2.17% +/- 0.04% (micro average: 2.17% +/- 1.65%)
squared_error: 5.702 +/- 0.133 (micro average: 5.702 +/- 8.068)
correlation: 0.632 +/- 0.009 (micro average: 0.632)
squared_correlation: 0.400 +/- 0.012 (micro average: 0.399)
```

Figure: 5-fold cross validation of GBT of model A3

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.552 +/- 0.083 (micro average: 2.554 +/- 0.000)
absolute_error: 2.029 +/- 0.063 (micro average: 2.029 +/- 1.551)
relative_error: 2.31% +/- 0.08% (micro average: 2.31% +/- 1.80%)
squared_error: 6.521 +/- 0.425 (micro average: 6.521 +/- 9.726)
correlation: 0.589 +/- 0.019 (micro average: 0.539)
squared_correlation: 0.347 +/- 0.023 (micro average: 0.291)
```

Figure: 5-fold cross-validation of GBT of Neural Network of model A3

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 2.507 +/- 0.053 (micro average: 2.507 +/- 0.000)
absolute_error: 1.975 +/- 0.043 (micro average: 1.975 +/- 1.545)
relative_error: 2.24% +/- 0.05% (micro average: 2.24% +/- 1.77%)
squared_error: 6.287 +/- 0.265 (micro average: 6.287 +/- 9.501)
correlation: 0.571 +/- 0.018 (micro average: 0.571)
squared_correlation: 0.327 +/- 0.021 (micro average: 0.326)
```

Figure: 5-fold cross-validation of GBT of Ensemble (Stacking) model A3

- In 5-fold cross-validation of GBT of model A1, A2 and A3, my best model is A1, followed by A3, and A2 respectively.
- The neural network performs better from the previous case but it is worse in performance as a comparison to GBT. I am thinking it is due to sampling size is 15000 or I have not tuned the parameters properly.
- I have used ensembles (i.e. Stacking) techniques to improve my result. I have put Gradient Boosted Trees in Stacking model learner. So, I think GBT will ignore all the decision makers on the left.
- I think model A3 with GBT perform best in terms of all measures without having 5-fold CV, neural net and stacking. By using CV and stacking my result is not improved much but worsens. Selecting this model A3 because it contains both structured and unstructured column, so we are not losing any information and error is also less.

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

Extension

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.607 +/- 0.000
absolute_error: 0.410 +/- 0.448
relative_error: 0.46% +/- 0.51%
squared_error: 0.369 +/- 0.839
correlation: 0.983
squared_correlation: 0.966
```

Figure: Grid Optimisation of GBT

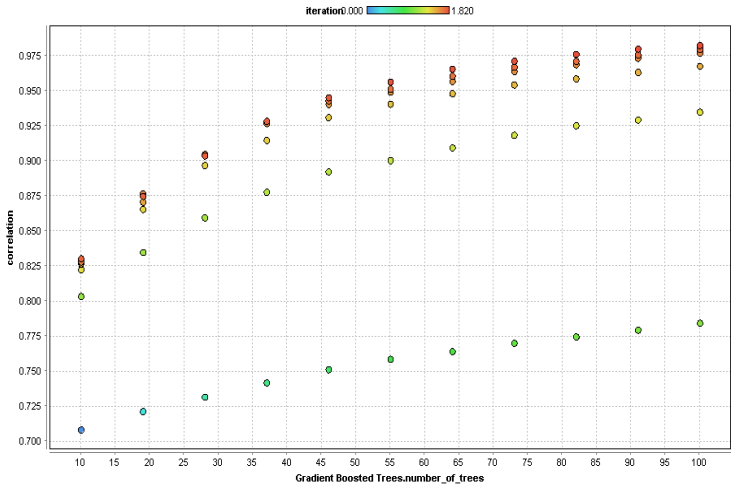


Figure: Correlation vs. Number of Trees

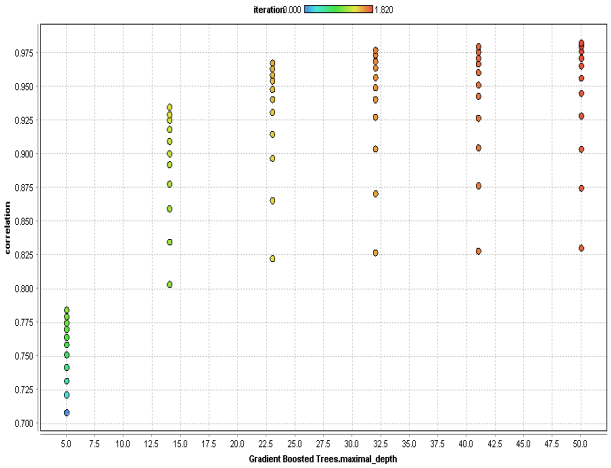


Figure: Correlation vs. Maximal Depth

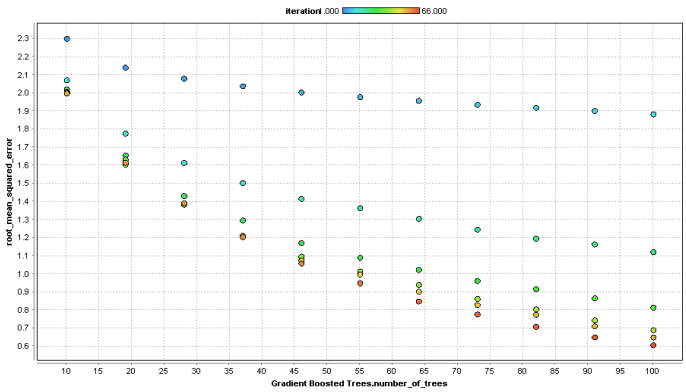


Figure: RMSE vs. Number of Trees

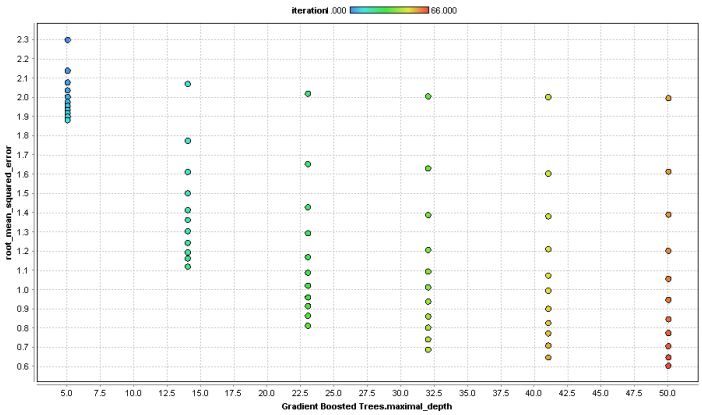
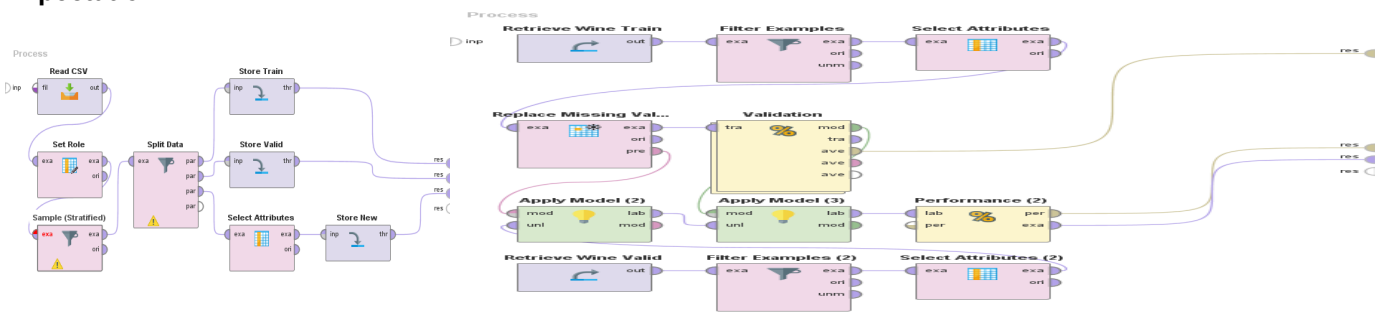


Figure: RMSE vs. Maximal Depth

- I changed booster parameters of GBT (A3) like a number of trees & maximal depth because to control over-fitting as higher depth and more trees will allow the model to learn relations very specific to a particular sample. I changed a number of trees from 10 to 100 in steps of 10 and maximal depth from 5 to 50 in steps of 5. So, there are 2 hyperparameters with 66 combinations selected. I did this analysis on model A3 because it contains both structured and unstructured column in it. So, we are not losing any essential data from the dataset.
- This is the best model so far I have. You can see from the performance vector all errors are small. For estimating points we need to use the GBT model with a number of trees should be 60 to 70 and maximal depth should be any value 30 to 40. All this parameter value depends on the time we have, application-related, and what percentage of error your application needs. All these values decide after taking these points of consideration.
- We can see as the number of trees increases the correlation also increases and RMSE value decreases. After 11th iteration correlation increases drastically. After that, it not changed much.
- As we increase the maximal depth the correlation increases and RMSE decreases. As we see from the graph when we have a maximal depth of 22, RMSE value is less and the correlation value is high.

Provide an Integrated Solution in Rapid Miner (one page)

Expectation



PerformanceVector

PerformanceVector:
root_mean_squared_error: 2.613 +/- 0.025 (micro average: 2.613 +/- 0.000)
absolute_error: 2.059 +/- 0.021 (micro average: 2.059 +/- 1.609)
relative_error: 2.33% +/- 0.03% (micro average: 2.33% +/- 1.85%)
correlation: 0.538 +/- 0.008 (micro average: 0.538)
squared_correlation: 0.290 +/- 0.008 (micro average: 0.289)

Figure: Test Result

PerformanceVector

PerformanceVector:
root_mean_squared_error: 2.645 +/- 0.000
absolute_error: 2.090 +/- 1.621
relative_error: 2.37% +/- 1.86%
correlation: 0.534
squared_correlation: 0.285

Figure: Validation Result

- Use stratified sampling to reduce your wine data to 5000 examples (use 2019 as random seed). Split data into three partitions 0.7, 0.2 and 0.1 for model training, validation and as new data. A new data will be lacking the label. Create a sample GBT predictive model with a mix of text and structured data.
- Retrieve the training data set. Filter out all missing label examples. Then Select attributes. Replace all missing values with average. Use unique integers to encode polynomial & text attributes. Z-normalise all numerical attributes. Create “GBT” and cross-validate it using “bootstrapping” (random seed = 2019). Measures the model performance with RMSE, absolute error, correlation. Run the process and see the result.
- The validation dataset performs similar to the testing dataset. There is not much difference between them. If new dataset comes in the future it will perform similar to the validation set.

Extension

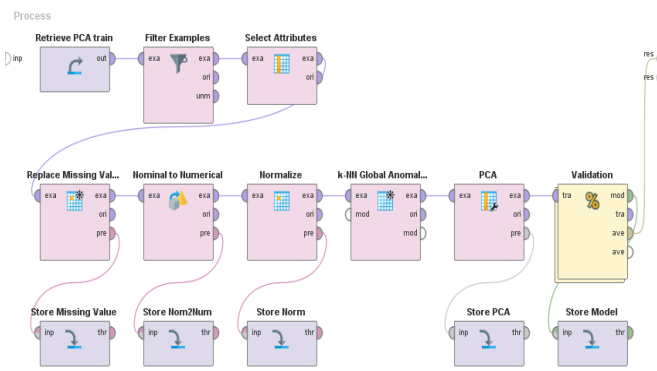


Figure: Store Process of PCA & k-NN

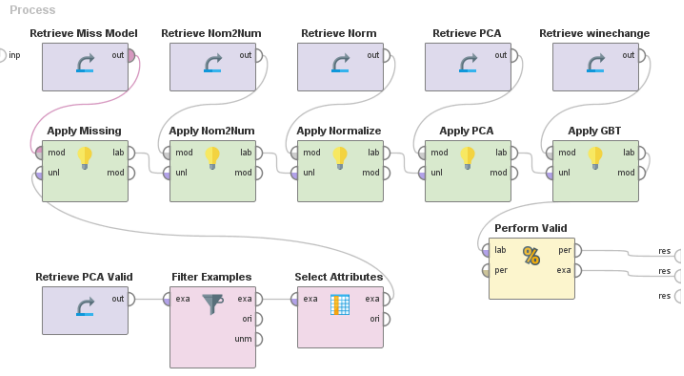


Figure: Retrieval Process of PCA & k-NN

PerformanceVector

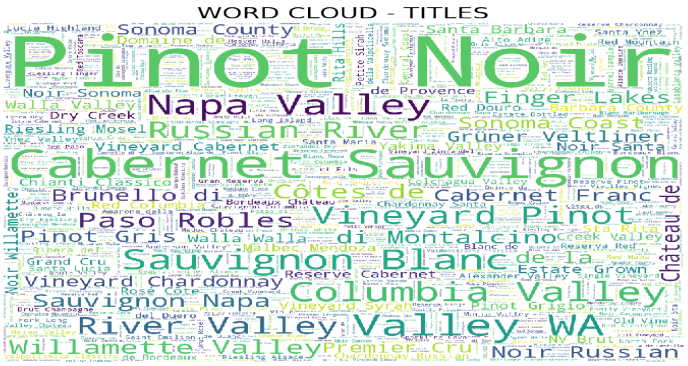
PerformanceVector:
root_mean_squared_error: 2.591 +/- 0.042 (micro average: 2.591 +/- 0.000)
absolute_error: 2.075 +/- 0.041 (micro average: 2.074 +/- 1.552)
relative_error: 2.35% +/- 0.05% (micro average: 2.35% +/- 1.79%)
correlation: 0.540 +/- 0.018 (micro average: 0.540)
squared_correlation: 0.292 +/- 0.019 (micro average: 0.291)

Figure: Performance of my GBT model (A3) after removal of outlier

- My model performance improved after removal of outliers.

Further Research and Extensions in RM (one page)

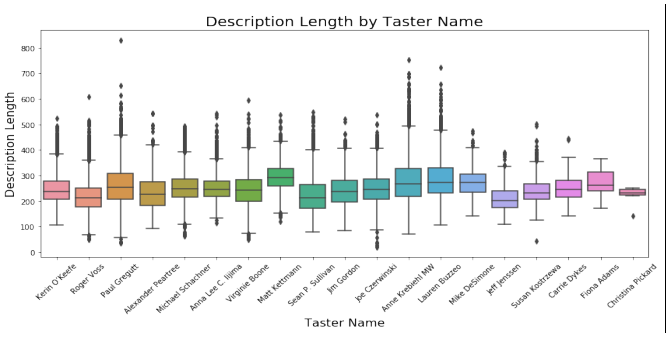
Expectation

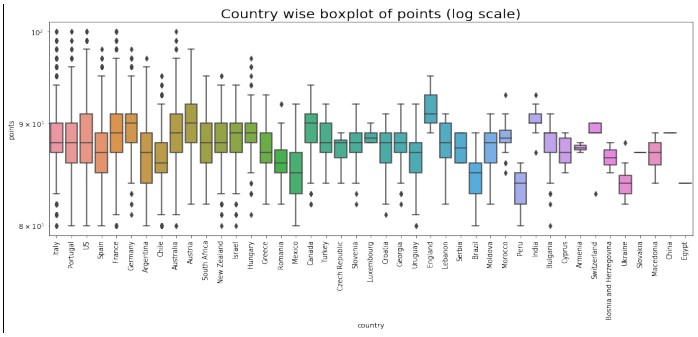


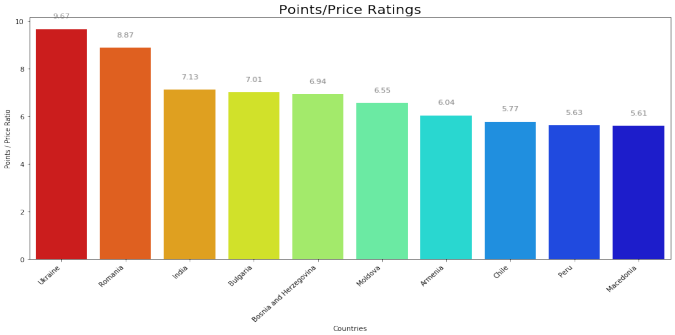
It seems like full-bodied, Cabernet Sauvignon, Pinot noir, and palate offer are the most mentioned characteristics, and Cabernet Sauvignon is the most popular of them all. This aligns with the fact that Cabernet Sauvignon "is one of the world's most widely recognized red wine grape varieties. Most of the review and title mentioned a lot about red berry, Colombia Valley, black cherry, fruit flavours, dried flavours and the aromas of the wine. Now let's take a closer look at the reviews and title for Australia.

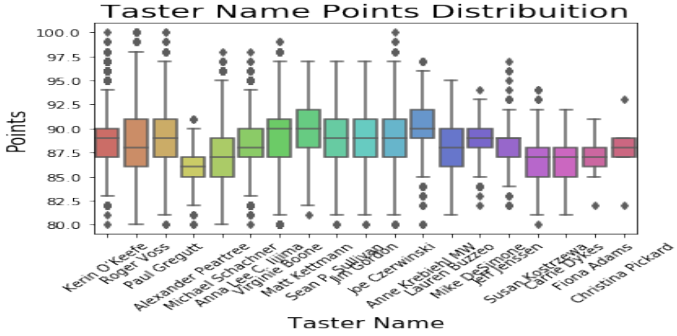
Commonly used words are to describe the Top Australian wines are Cabernet Sauvignon, Shiraz McLaren, McLaren Vale, full-bodied, medium bodied is the most mentioned characteristics. South Australian Wines are mentioned more comparison to other states.

Extension









Wines with highest points also have the biggest descriptions length. The wine produced in France, Australia, Portugal, Italy, the US receives top rating while wine in Uruguay, South Africa, and Portugal revives the least rating. Based on the points/price ratio analysis by country, we conclude that Ukraine is the best country followed by Romania and India.