Investing in manufacturing robotic household appliances in South Africa

Emerald Adams
4133428 University of the Western Cape

Contents

Organisational Understanding

The business would like to know if there is a potential investment opportunity in South Africa particularly in the manufacturing sector for robotic household appliances. The aim is understand if it is feasible and viable to manufacture these products in South Africa.

The business seeks to make an informed decision about investing in South Africa by analysing the manufacturing industry. The business will investigate the employment rate in different provinces within the manufacturing industry, household appliances manufacturing sales and manufacturing's contribution to the Gross domestic product.

The focus of this investigation is South Africa as the target market for manufacturing robotic household appliances. The analysis will focus on current manufacturing data and anticipation of future trends. The international company will make the consideration for this investment based on the findings on this investigation.

Data sources that will be included is information on the manufacturing performs with the GDP (Gross domestic Product), employment statistics within the manufacturing sector and how household appliances performance in terms of manufacturing.

The collection of the data will allow for comprehensive assessment of the manufacturing landscape in South Africa allowing for in-depth analysis of key factors influencing investment decisions. This data is collected through government reports and statics. A correlation matrix will be used to interpret the data about employment in different provinces in the manufacturing industry. This will show which province it is best to invest in to manufacture household robotic appliances. Linear regression will be used to analyse how the manufacturing industry performs relative to the Gross Domestic Product. Finally, the Gradient Boosting Model will be used to predict how household appliances perform and use historical data to predict how they will later perform.

Data Understanding

The Data used is made up of three different data sets, but the data is collected from the same source. The data has been collected from the stats sa website which is provided by the Department of Statistics South Africa

The different data sets have been collected at different times but they there are all updated to the date 2023. The data for the employment statistics is Labour force data which has been recorded from 2008 to 2023. This data set is. Excel format has many different sheets in the Excel file, but the focus will only be Table 3.2 which describes the different sectors in manufacturing. The data set for the manufacturing sales and production in South Africa is recorded from 1998 to 2023. The focus of that dataset will be how household appliances perform. The last data set is how the manufacturing industry affects the GDP. This data is recorded from 2017 to 2023. This will show if the manufacturing sector is a positive contributor to the GDP.

All the datasets are linked to the manufacturing sector. The datasets will give insights into how the manufacturing industry performs and how household appliances perform. The importance of household appliance performance is closely linked to how the robotic household appliances will perform as there is no market like it in South Africa now.

The data is publicly available at the Stats SA website which means that it is accessible to anyone interested in analysing it. The datasets are from the Department of Statistics South Africa which means it can be considered legitimate and secure.

The data is recorded up to 2023, showing the data is relevant to today. The data is from the department of south Africa and because of tis it can be recognised as legitimate.

The dataset names:

1. Quarterly Labour Force Survey: (Table 3.2)
2. Manufacturing Product and Sale: (Excel data from 1998)
3. GDP P0441:

CRISP-DM Cycle 1: Classification Model

Correlation model

The correlation will used for the employment in different provinces in the manufacturing sector because it is able to identify patterns or relationships between variables. (North, Correlation Methods, 2018).By looking at the total employment then using the correlation model to compare the employment in the different provinces it will be possible to identify which provinces has the highest employment in manufacturing. This will identify where investors should manufacture their household robotic appliances.

Data Preparation

1. Import the dataset into RapidMiner using the Read Excel operator. Run the model and then go result view and then analyse data to check that the data type is integer and that there are no outliers.

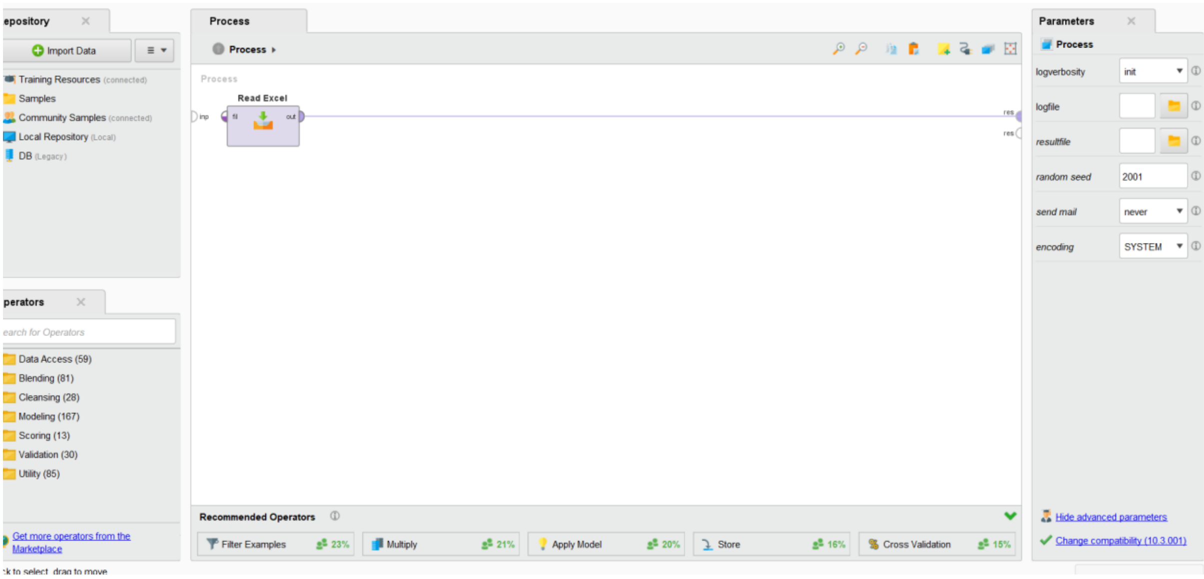


Figure 1:import data for correlation model

2. Add a set role operator and make the date attribute as an id, so that it can be exclude the from the correlation model. The date attribute would skew our output and it is not an integer so it needs to be excluded.

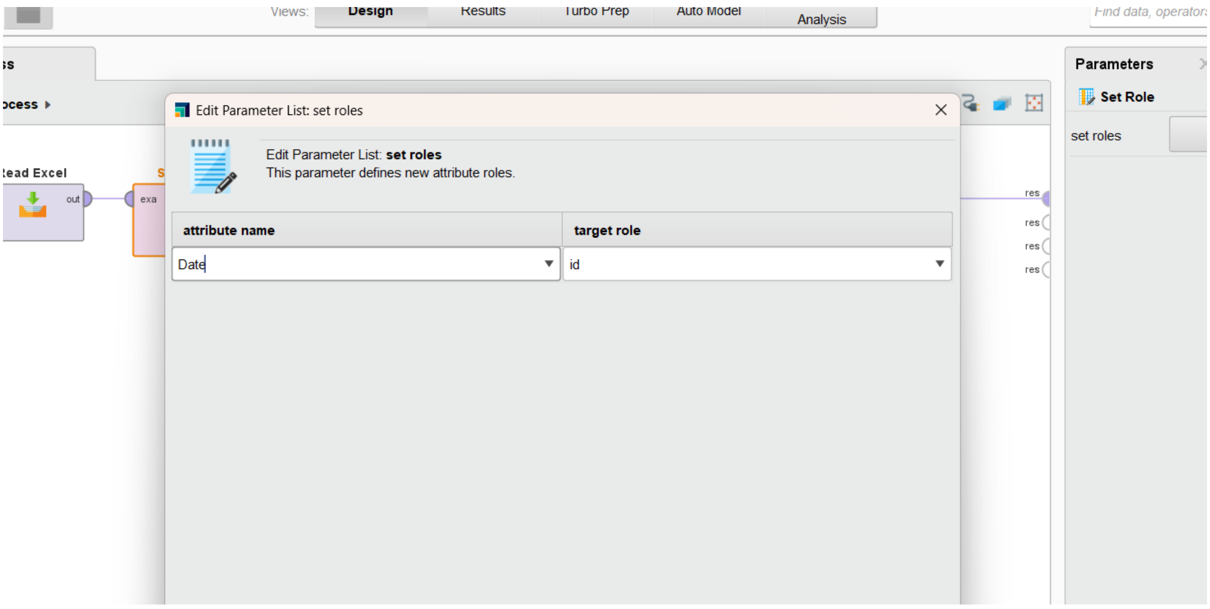


Figure 2: Set role for correlation model

Modelling

3. Add the correlation matrix operator. Connect the exa, wei and mat port. As in the figure below

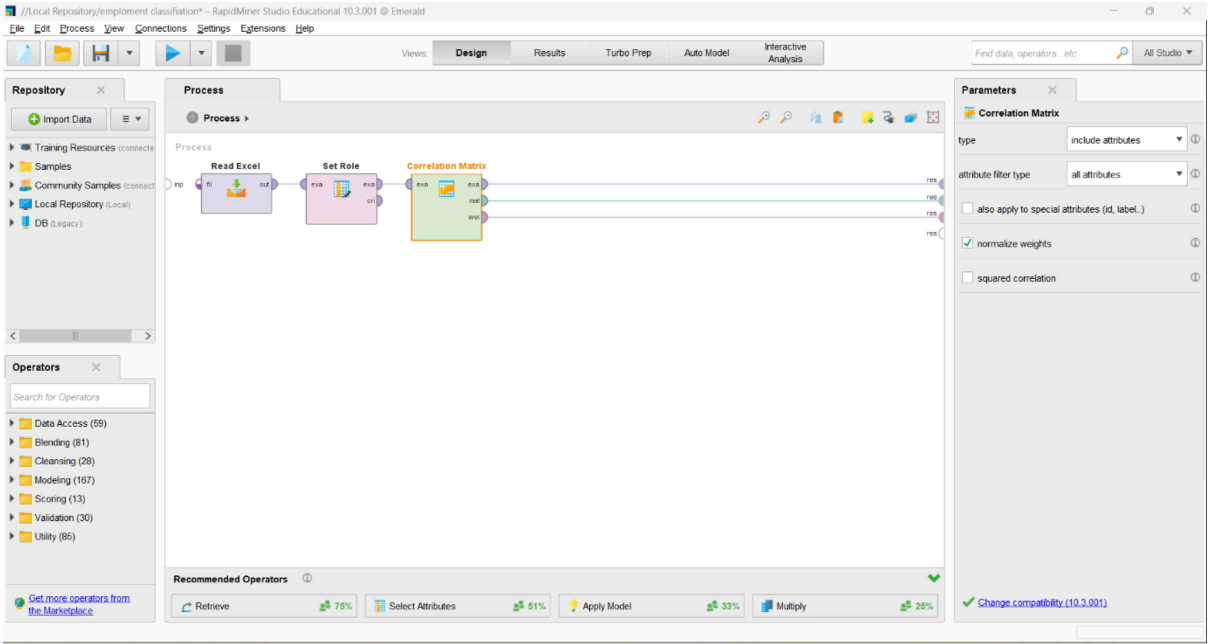
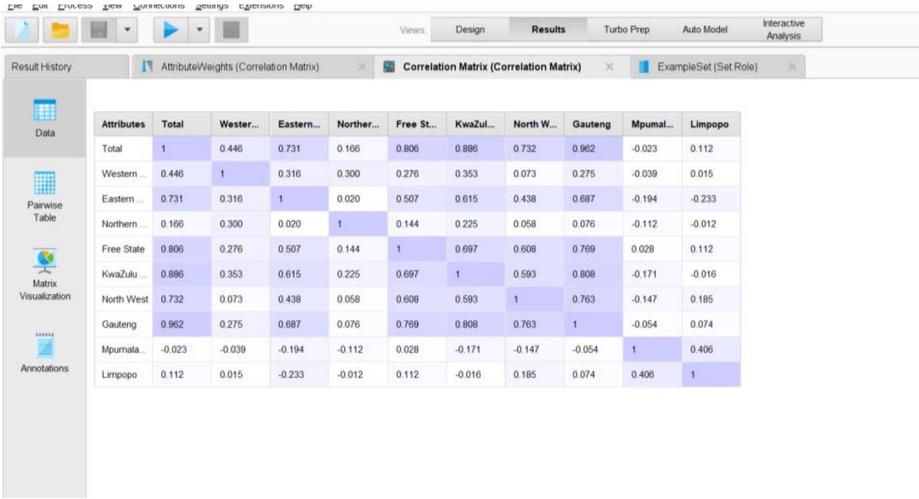


Figure 3:Correlation matrix operator

Evaluation



Attributes	Total	Wester...	Eastern...	Norther...	Free St...	KwaZul...	North W...	Gauteng	Mpumal...	Limpopo
Total	1	0.446	0.731	0.166	0.806	0.886	0.732	0.962	-0.023	0.112
Western ...	0.446	1	0.316	0.300	0.276	0.353	0.073	0.275	-0.039	0.015
Eastern ...	0.731	0.316	1	0.020	0.507	0.615	0.438	0.687	-0.194	-0.233
Northern ...	0.166	0.300	0.020	1	0.144	0.225	0.058	0.076	-0.112	-0.012
Free State	0.806	0.276	0.507	0.144	1	0.697	0.608	0.769	0.028	0.112
KwaZulu ...	0.886	0.353	0.615	0.225	0.697	1	0.593	0.808	-0.171	-0.016
North West	0.732	0.073	0.438	0.058	0.608	0.593	1	0.763	-0.147	0.185
Gauteng	0.962	0.275	0.687	0.076	0.769	0.808	0.763	1	-0.054	0.074
Mpumala...	-0.023	-0.039	-0.194	-0.112	0.028	-0.171	-0.147	-0.054	1	0.406
Limpopo	0.112	0.015	-0.233	-0.012	0.112	-0.016	0.185	0.074	0.406	1

Figure 4:Correlation matrix.

Correlation coefficients are between -1 and 1. Positive correlation is between 0 and 1 and negative is between 0 and -1. The correlation matrix above shows the relationship between the total number of employed people in manufacturing and the people employed in manufacturing in different provinces. It is important to note just because there is correlation between variables doesn't mean there is a meaningful relationship between them such as relationship between different provinces doesn't give any meaningful insights. The relationship between the same attributes will always be 1. (North, Correlation Methods, 2018)

The relationships between the total number of employees in manufacturing and employment in the different provinces are meaningful. This will help identify which provinces have the most manufacturing employees. The best provinces to manufacture robotic household appliances based on the amount of skilled labour would be the Eastern Cape with a positive correlation of 0,731, the Free State with 0.866 correlation, KwaZulu Natal with 0.886, North West with 0,732 and Gauteng with 0.962. Overall, the most beneficial province to manufacture robotic technology would be Gauteng because it has the highest positive correlation. This means that if the manufacturing sector employment was to increase the employees in these sectors would also increase.

Deployment

The correlation matrix identified that Gauteng and KwaZulu-Natal have the highest positive correlation. This means that from the data if the total number of manufacturing employees were to increase it is likely that the employment in these industries would also increase. It is important to note that correlation does not always equal causation. Therefore, just because Mpumalanga has a negative correlation it doesn't mean that if the employment in the manufacturing sector was to increase that Mpumalanga employment in the manufacture sector was to decrease. Also, correlation is not percentages, it solely measures the strength of the relationship.

To consider if the business were to invest in the manufacturing sector in South Africa:

- Kwa-Zulu Natal and Gauteng has the largest manufacturing sector so I would be wise to invest in theses to industries. There is also existing infrastructure in these provinces. Also, Kwa-Zulu Natal has already third of manufactured exports being produced. (KZN Top Business Porfolio, 2024)

CRISP-DM Cycle 2: Prediction Model

Gradient Boosting

Gradient boosting model will be used to forecast household appliances manufacturing sales. Gradient boosting is known for its high predictive accuracy. The model is complex, but it still provides model summaries which is makes it easier to interpret. The model will enable investors to identify the sales performance, and such make informed investment decision using reliable insights from the gradient boosting mod (Geeks for Geeks, 2024) el.

Data Preparation

1. Add the data set into the Process window in RapidMiner. Connect to the out port to the res port as shown and then run the model.

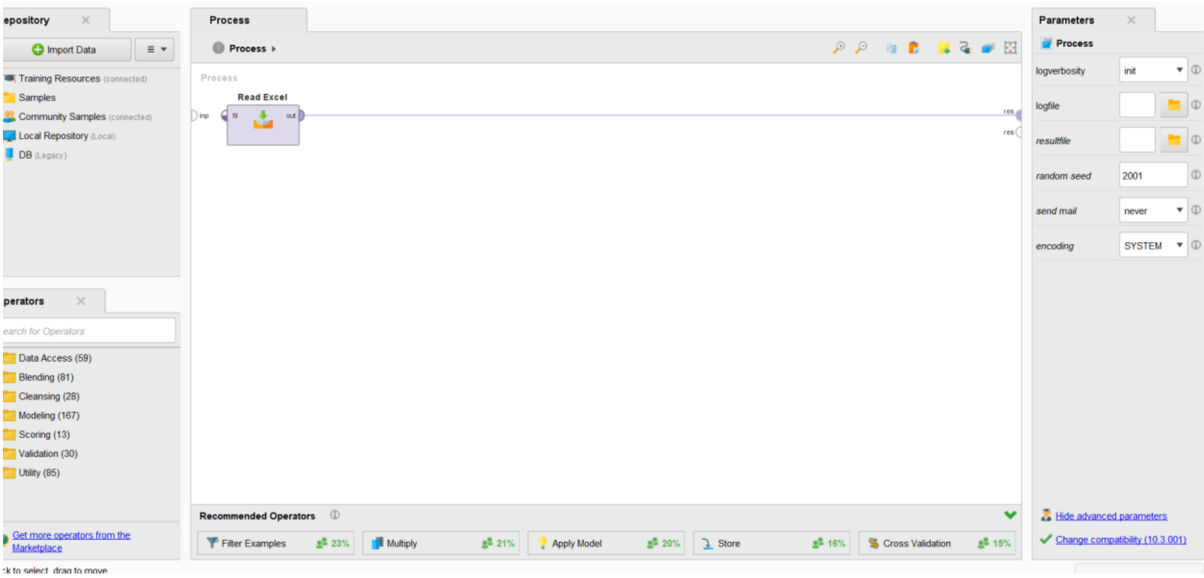


Figure 5: Import data for Gradient Boosting Tree

2. Go the result view and analyse the data. The data has outliers, so Normalisation is needed to prevent theses outliers. This will prevent outliers from skewing the model.

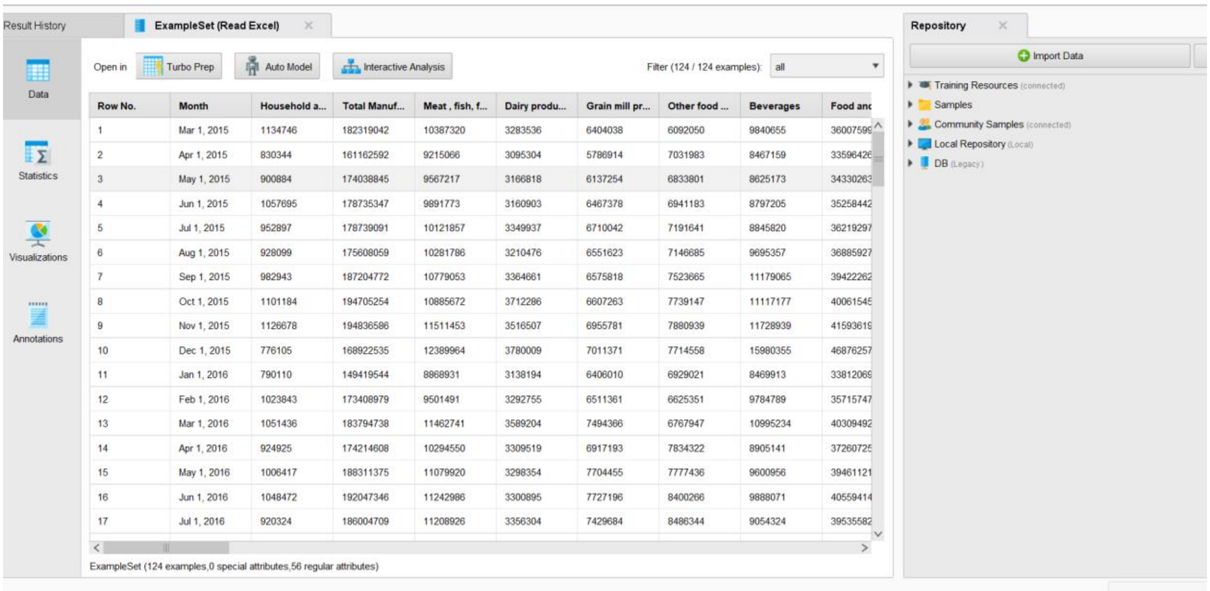


Figure 6: Analyse data for outliers.

3. Add the normalisation operator and make the range between 0 and 1. The min being 0 and the max being 1. Connect the exa port to the res port as seen below.

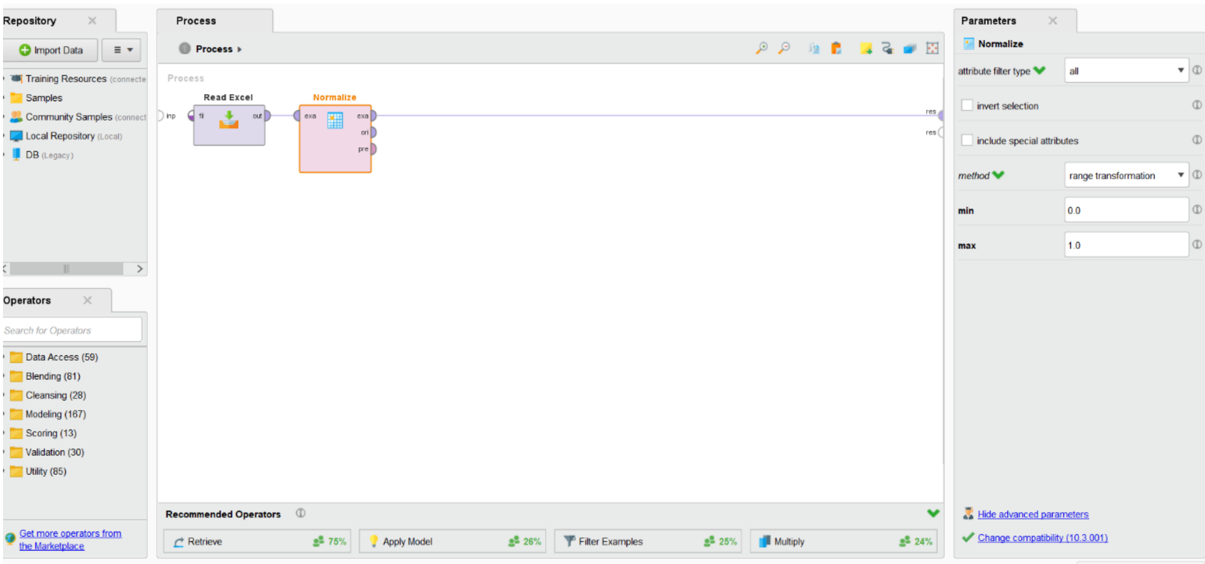


Figure 7: Normalisation operator for gradient boosting model.

4. There are missing values present, to filter out the missing values use the Filter Example operator. Configure the operator to invert the selection and choose the attributes with missing values. Connect the output of this operator to the next step.

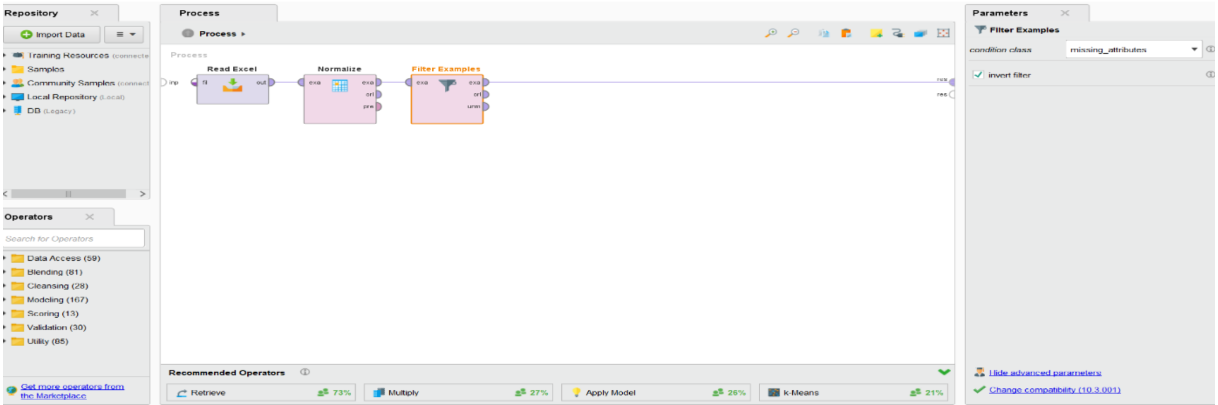


Figure 8: Remove missing values.

5. Windowing is used for time series data to create windows of data for training the model. Add a Windowing operator to define the time windows for your data. Configure the operator to specify the size and step of the windows. Connect the output of this operator to the next step.

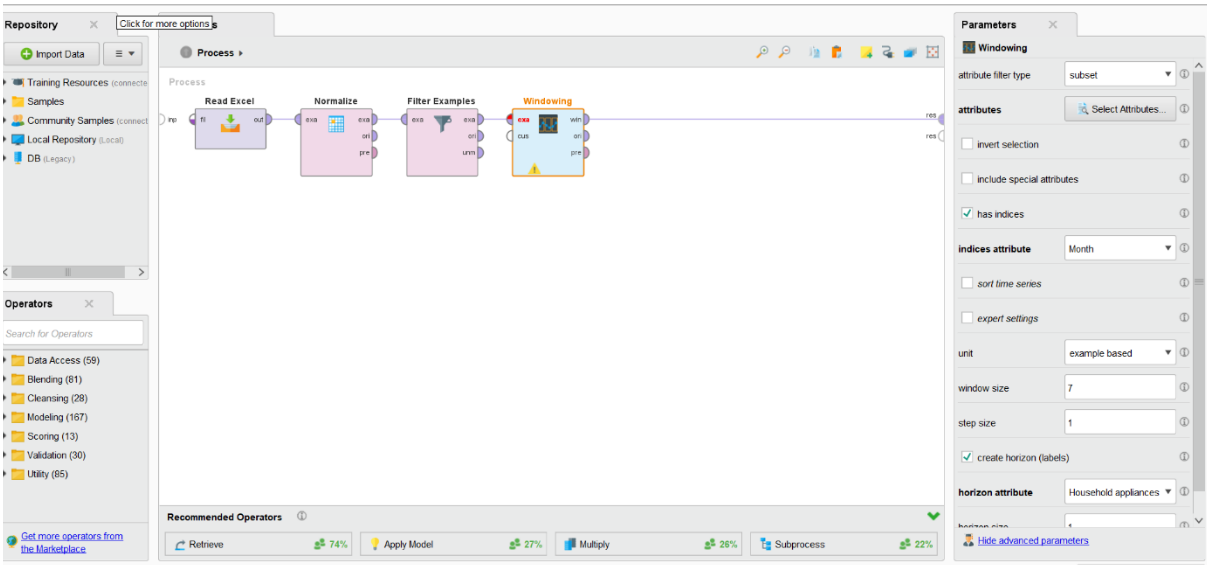


Figure 9:Windowing operator

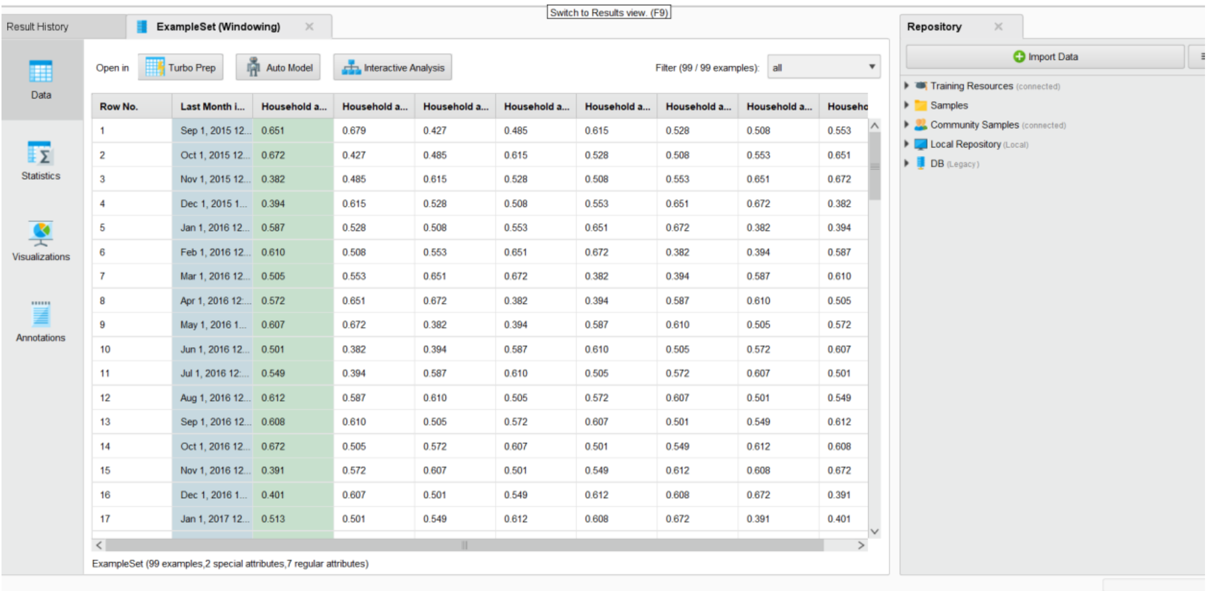


Figure 10: Windowing operator output.

6. Now add the sub process operator

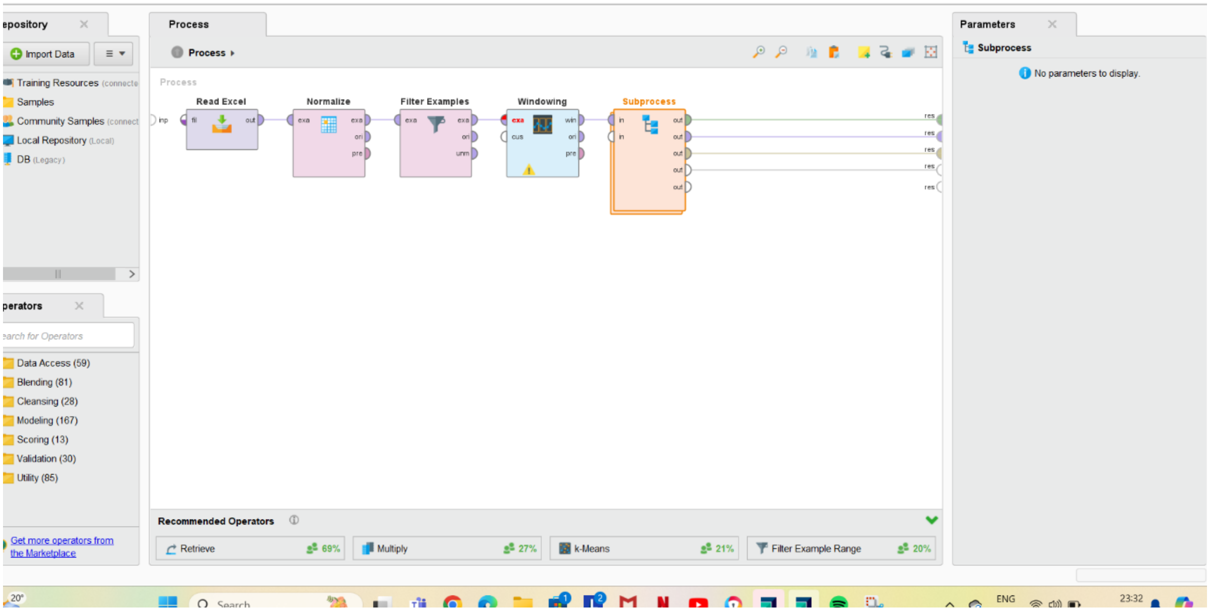


Figure 11: Add subprocess operator.

7. Now double click on sub process. To ensure robustness of the model, set up cross-validation. In the subprocess, add a generate macro attribute and add a cross-validation operator. Connect the generate macro-operator to the “in” port and then the add the cross validator to the generate macro-operator tusing the “thr” port. Add a cross validation operator and store operator. Connect the Store operator to “mod” port of the cross-validation operator and the Store (2) operator to “tes” port of the cross-validation operator and Store 3 operator to “per” port of the cross-validation operator and connect all the store operator to “out” ports.

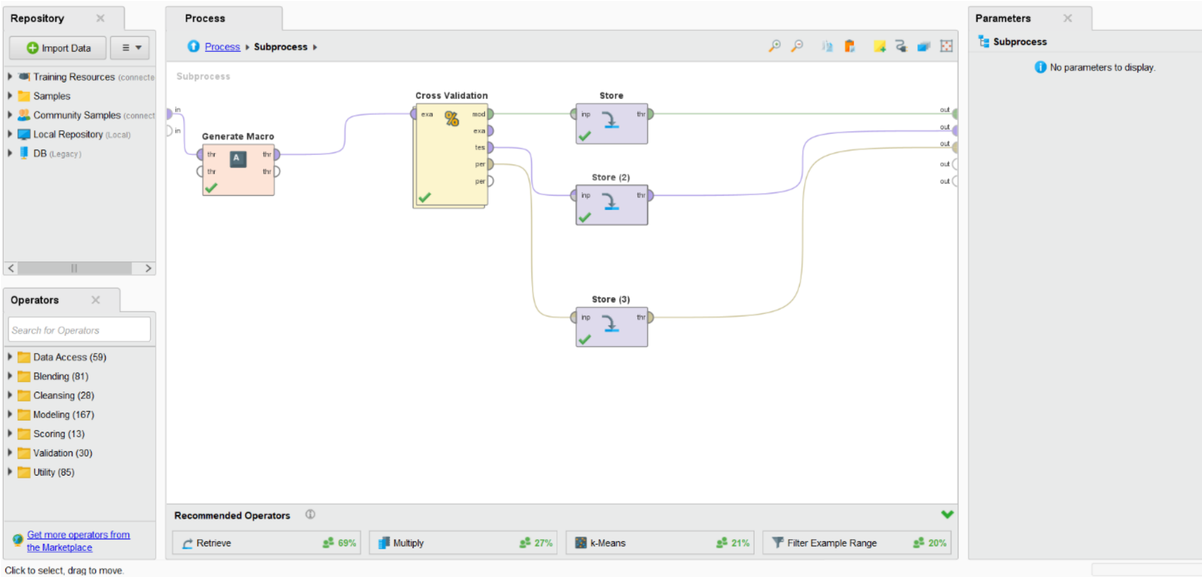


Figure 12: Setting up the cross-validation operator.

8. Now in the generate macro-operator in the Edit parameter list change macro name to “model training” and the function expressions to “Windowing”. Thereafter double click on cross validation to start the modelling process.

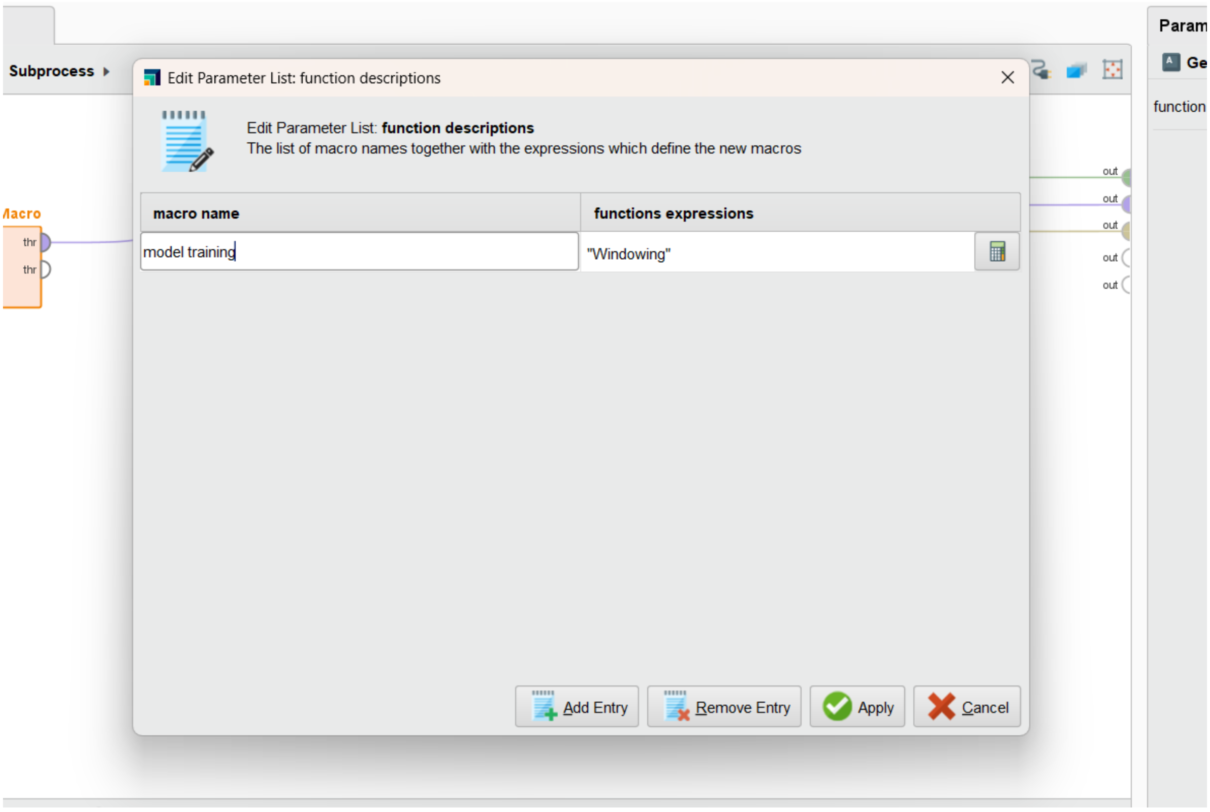


Figure 13: Generate Macro parameter.

Modelling

9. Use the search operator and add gradient boosting operator to the training window and connect it to the “tra” port and then connect gradient boosted tree operator to the “mod” port and the “exa” port to the “mod” port and “thr” port. Now, once the model is trained in the testing windowing add the apply model operator and connect to the ports below.

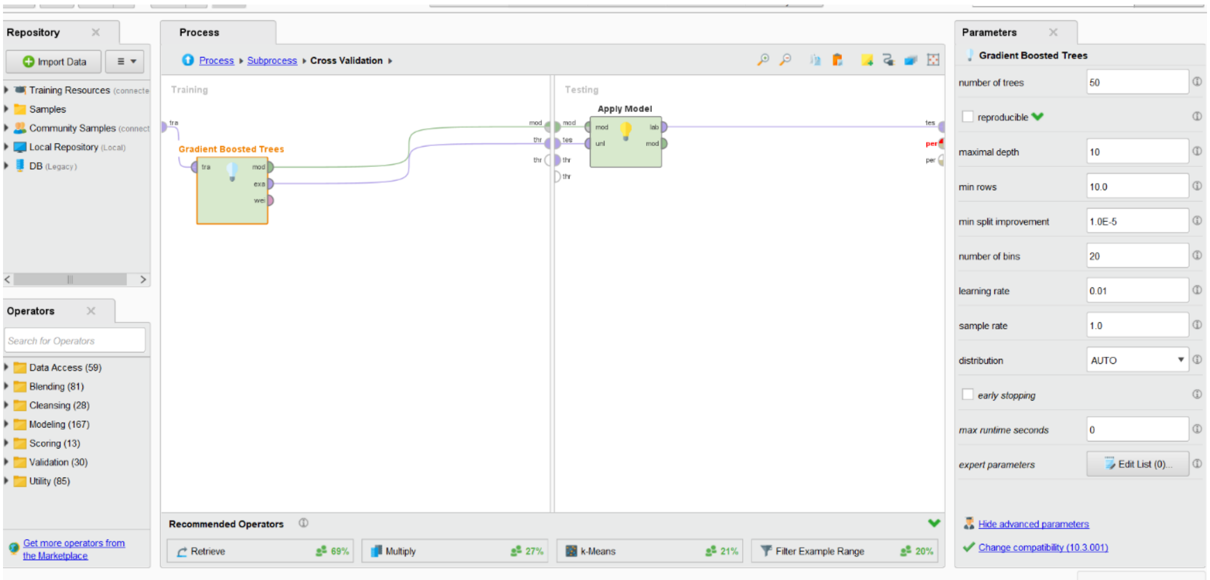


Figure 14: Gradient Boosting operator and Apply model operator

10. Now the next step is to add a performance metric for regression operator. By the parameters choose absolute error and relative error. And now run the model.

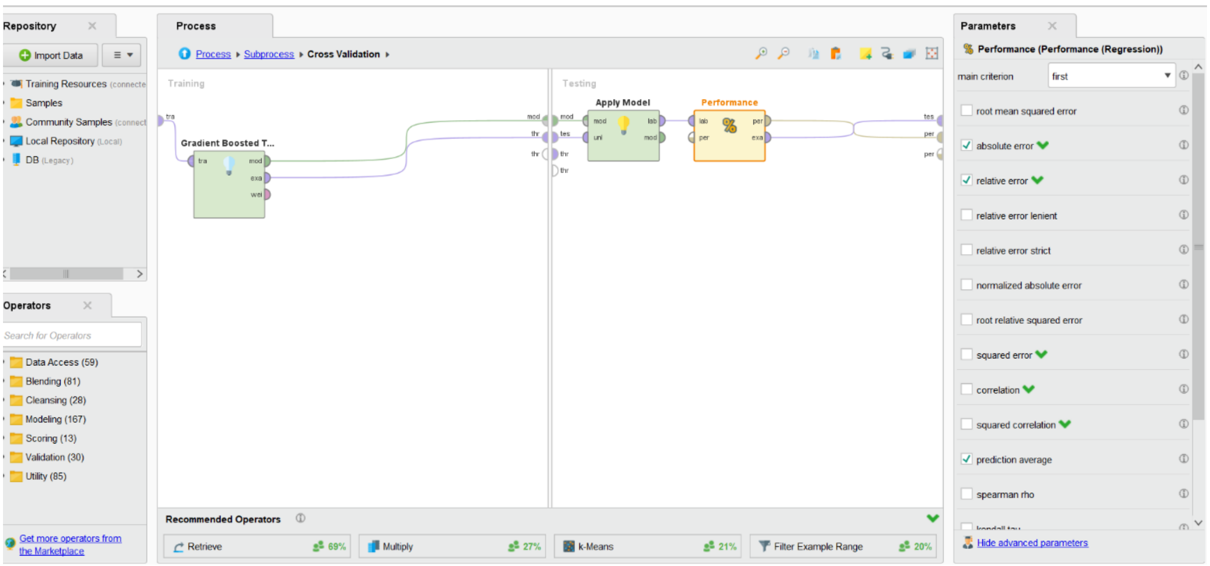


Figure 15:Performance metric operator.

Evaluation

Gradient boosting is a prediction model that combines the predictions from several models to improve the overall predictive accuracy. The term gradient refers to the method using the gradient of the loss function to minimise errors during training stage of the model and the word boosting refers to the model combing weak predictive models to form a strong learner. (Geeks for Geeks, 2024)The model used is Gradient boosting for time series forecasting in the context of household appliances manufactured and sold in South Africa. The model is used to predict future demand or production levels based on the historical data.

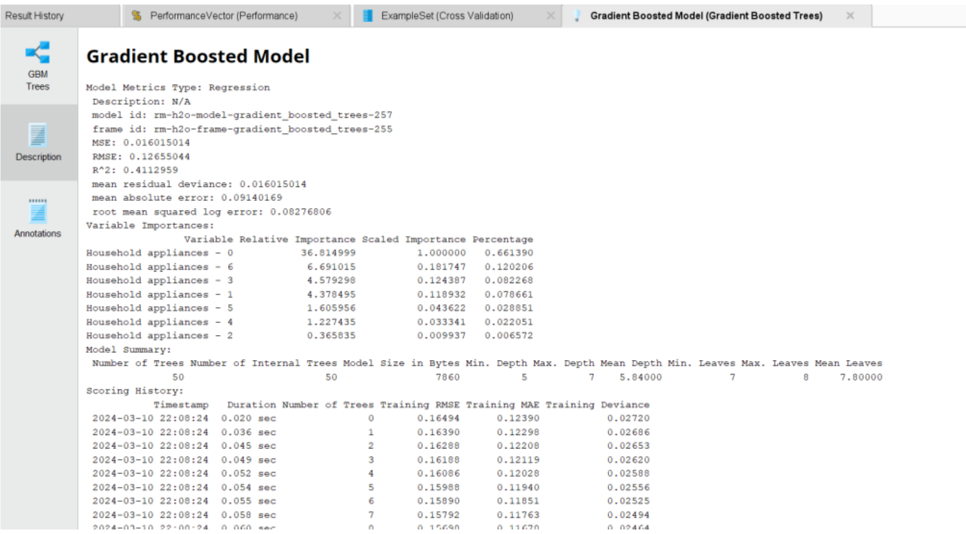


Figure 16: Gradient boosting metrics.

The gradient boosting model has the following metrics.

1. Mean Squared Error (MSE) measures the average squared difference between the actual; and predicted household appliances manufacturing sale values. The MSE above is 0.0016015014 which indicates that the model's prediction has a relatively low squared error on average.
2. Root Mean Squared Error (RMSE) is the square root of the MSE and measures the average deviation of the predicted values from the actual values. The RSME of the gradient boosting model is 0.12655044 meaning the average the predictions are off by is approximately 0.12655044 units.
3. R-Square represents the proportion of variance in the dependant variable tat is predicted from the independent variable. The R-squares is 0.4112959 meaning that approximately 41,13% of the variance in the dependant variable (household appliances manufacturing sales pricing).
4. Mean Absolute Error (MAE) is the measure of the average absolute difference between the predicted and actual values. The MAE is 0.9140169 which indicates that the predictions are off by 0.09140169.
5. Root Mean Squared Log Error (RMSLE) measure the ration between the prediction and the actual value. The RMSLE is 0.08276806 which indicates the average ratio between the predicted and actual values.
6. Variable importance shows the importance of each feature in making a prediction, the Household appliances – 0 has the highest importance.
7. The model summary explains the complexity of the model such as number of trees, which is 50, model size which is 50 and depth which is 10 and the number of trees which is about 7,8.

The Gradient boosting model has a decent performance with a relatively low error metrics and a reasonable R-squared.

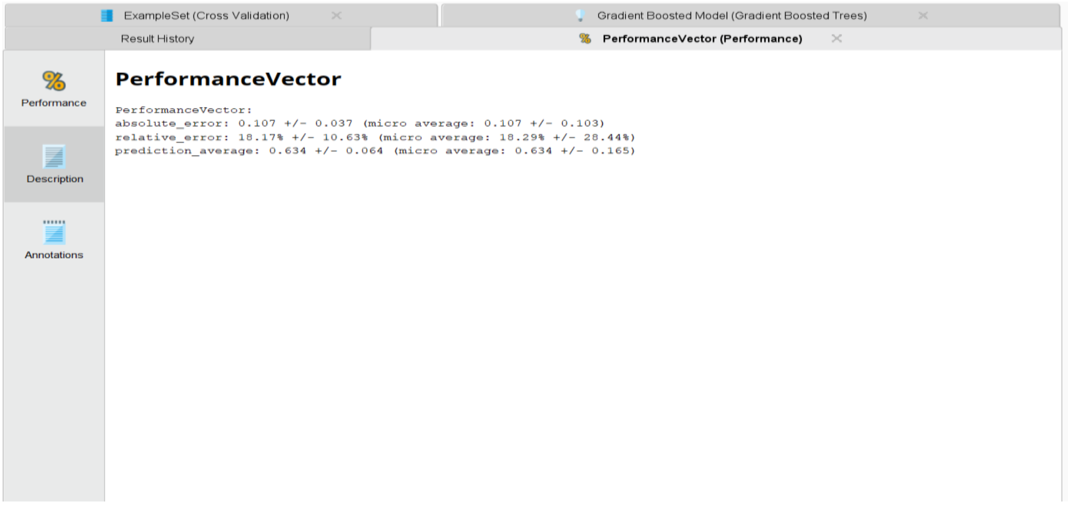


Figure 17: Performance metrics.

The performance metric of the gradient boosting model, we have the is following. Th absolute error measures the average difference between the predicted values and actual values. The value is 0.107, with the standard deviation of 0.037. This indicates that models' prediction deviates from the actual values by around 0.107 units with some variance around the average. The micro average shows similar results but with a higher standard deviation of 1.103, which indicates more variability across different instances. The relative error measures the percentage difference between the predicted actual values by 10.63%. The is indicates that on average the model's prediction deviates from the actual value by around 18.17% with some variability around this average. The micro average indicates a similar relative error but a higher standard deviation of 28,44%. The average prediction of the model is 0,634 with a standard deviation of 0.064. The micro average shows a similar average of 0.35 but with a higher standard deviation.

Deployment

The model's ability to make accurate predictions about future demand or production levels of household appliances is quite good based on the relatively low error metrics MSE, RMSE, MAE, RMSLE and R-squared values The model has standard deviation of 0.037 which indicates prediction accuracy across instances. It is necessary to identify and address sources of variability to improve the model performances. The average prediction of household appliances sales after manufacturing is 0.63, the data has been normalised thus to find the accurate sales value, it will be necessary to denormalize the results. The data range is between 300000- 1000000 and the average being 0.63 the average prediction is more than the average range.

The gradient-boosting model can be used to predict how household appliance manufacturing sales are to perform in the future. Continuous monitoring and evaluation of the model's performance to make sure it stays accurate and reliable over time is necessary. An ongoing validation and refinement based on real-world feedback and changing market dynamics will be necessary,

Conclusion

The evaluation of the gradient boosting model for time series forecasting of household appliances in manufacturing sales in South Africa is effective based on the performance metrics. The average prediction of household appliances sales is 0.63. This indicates its utility informing production planning and resources. It is essential to denomalise the result to get an accurate amount of the sales. However the average amount is above half the range of the results, which is promising. The Deployment of the gradient boosting model offers insights about the future sales trends and allows for decision making about whether investors should manufacture the products in South Africa.

Linear Regression

Data Preparation

1. Import Data set with Read excel operator. Connect the out ports to res ports. Analyse the data, check the variable data types, The variable data types are Real, for linear regression to work it needs integer variable types. Add a Real to integer data type to change the variables from real to integer. The next step is to add the set role operator. After adding the set role operator, in the edit list of the set role operator make the industry attribute the ID because it is a date and the GDP at market prices the label as this is the target variable.

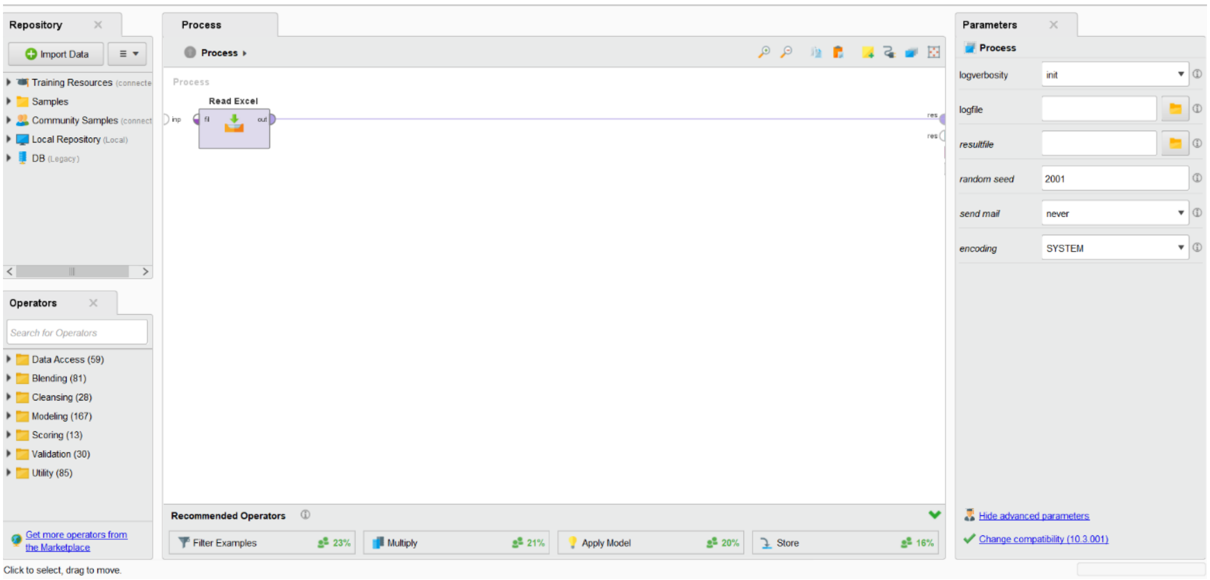


Figure 18 Import data for linear regression model

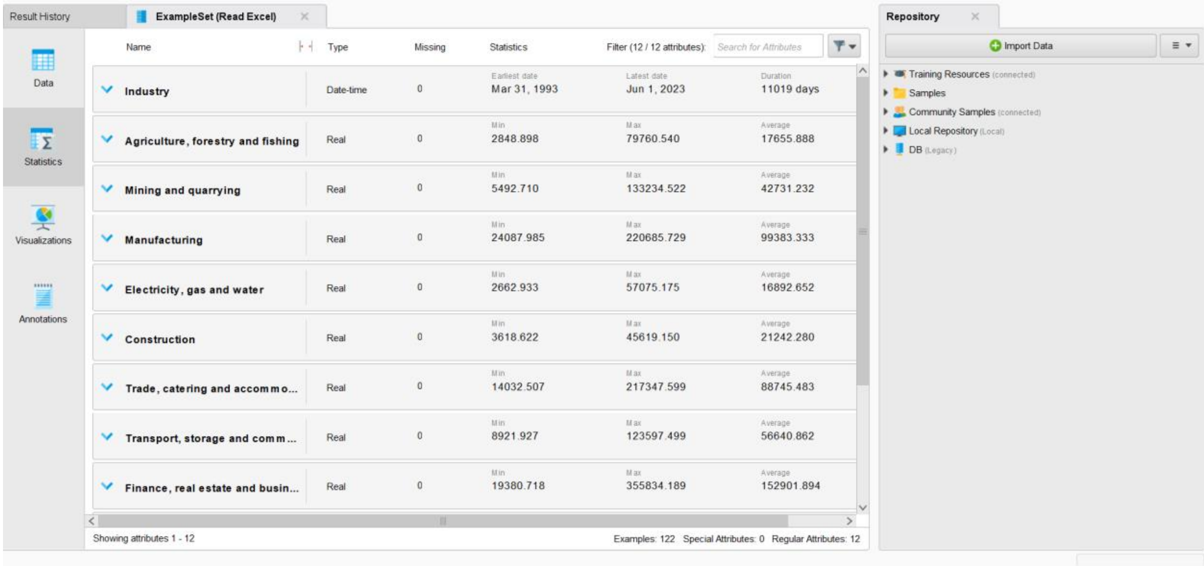


Figure 19: Analyse linear regression imported data

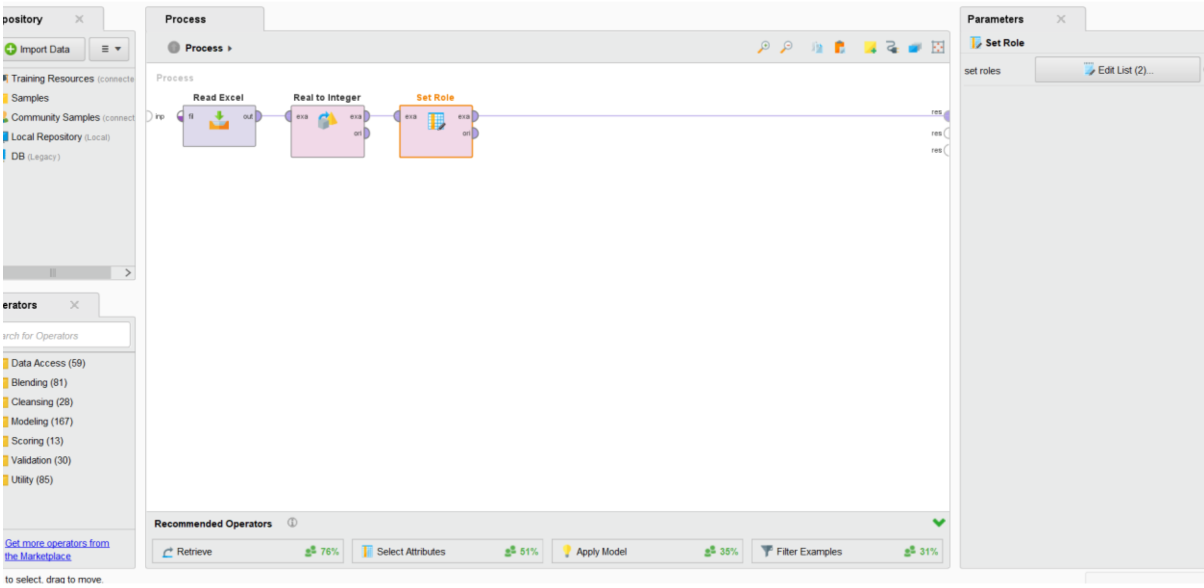


Figure 20: Add a set role for linear regression model

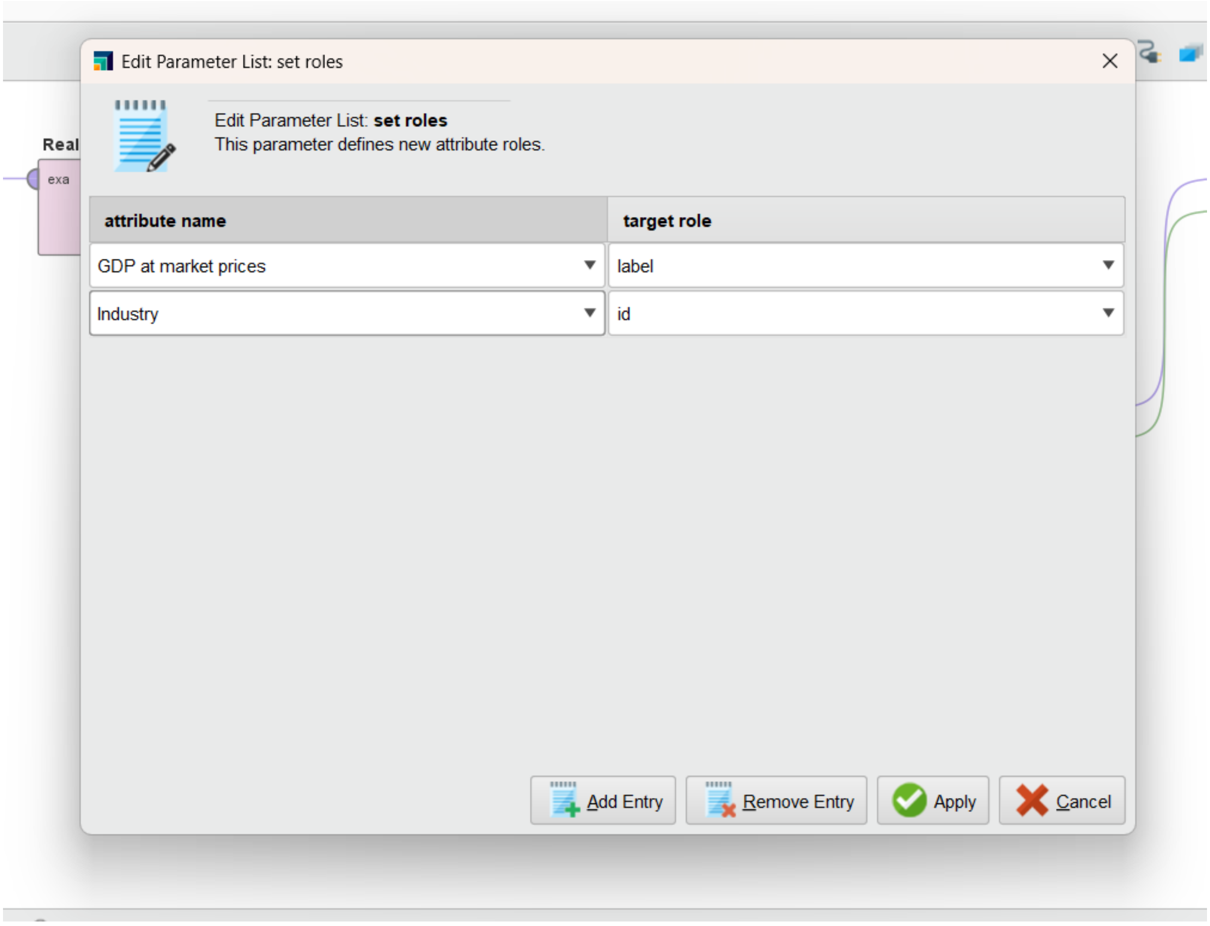


Figure 21: Add label and ID

2. Next use the split data operator because linear regression needs a training and scoring data set. The ratio of the split should be 0.7 and 0.3. This should be done in the split data operator parameter list.

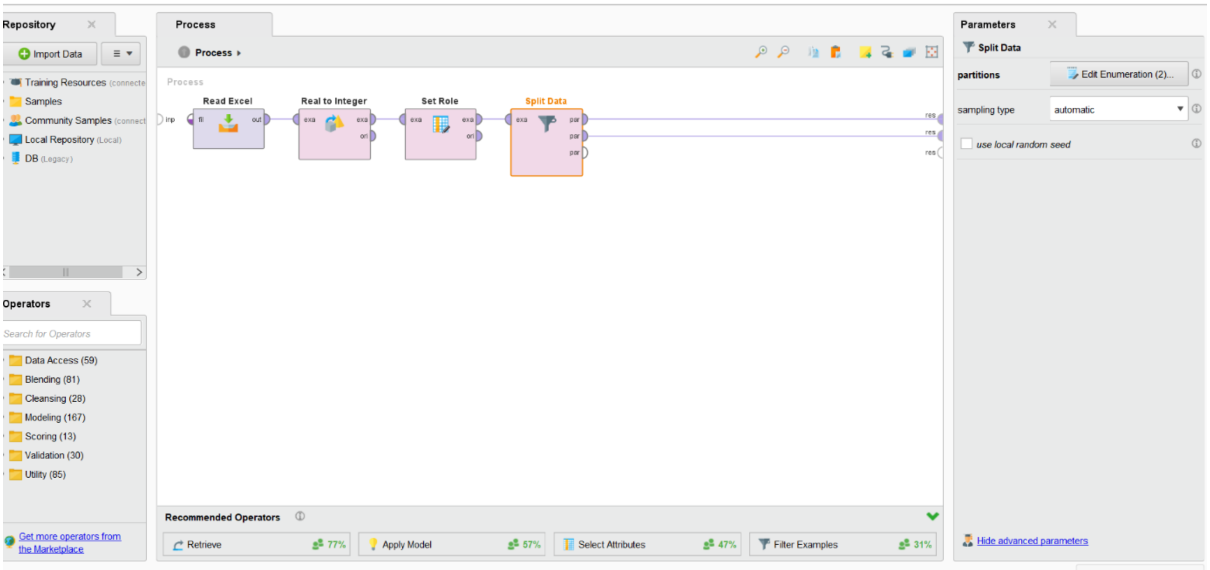


Figure 22:Add a split data operator

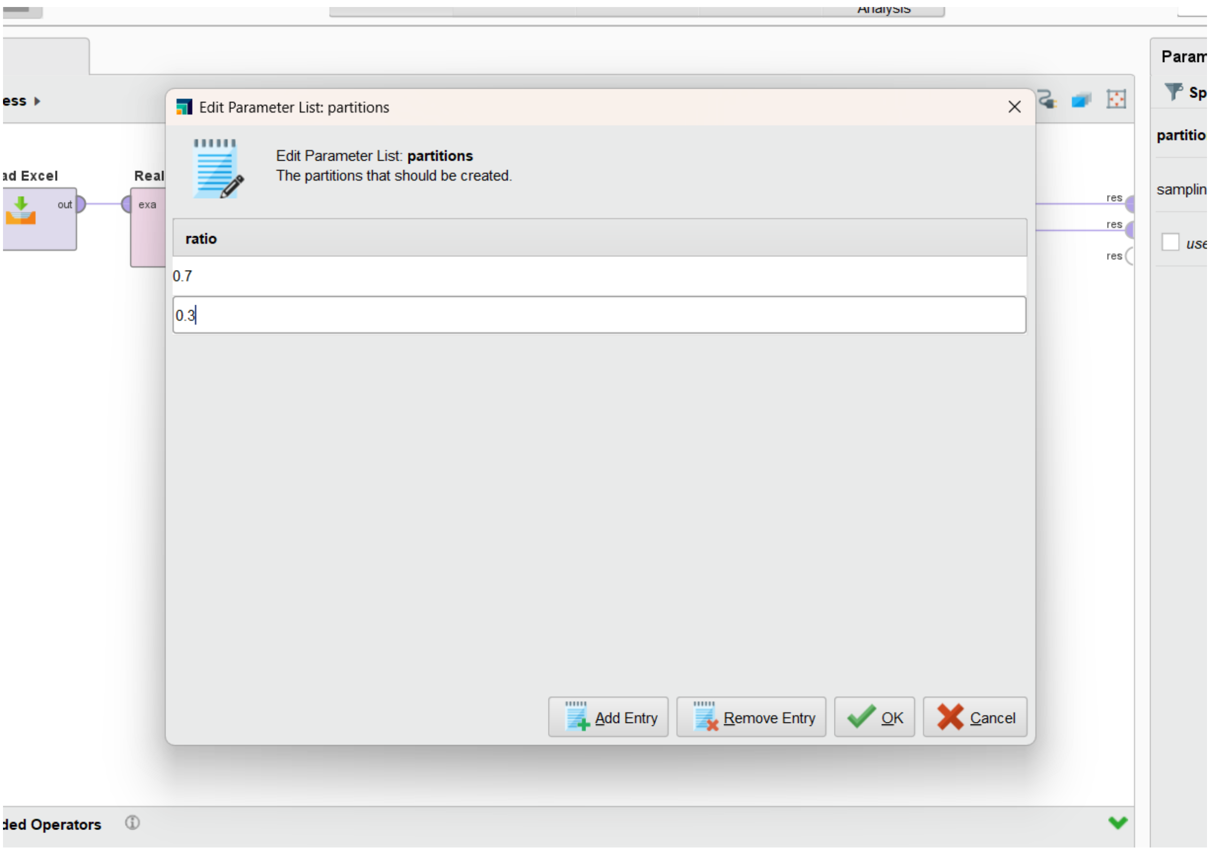


Figure 23: Ratio of split data operator parameter

Modelling

3. Add the Linear regression operator in the design view. The min tolerance should be 0.05 which is also known as the confidence level and the ridge should be 1.0E-8. The min tolerance of 0.5 is the common statistical analysis thus this is the default. (North, Linear Regression, 2018)

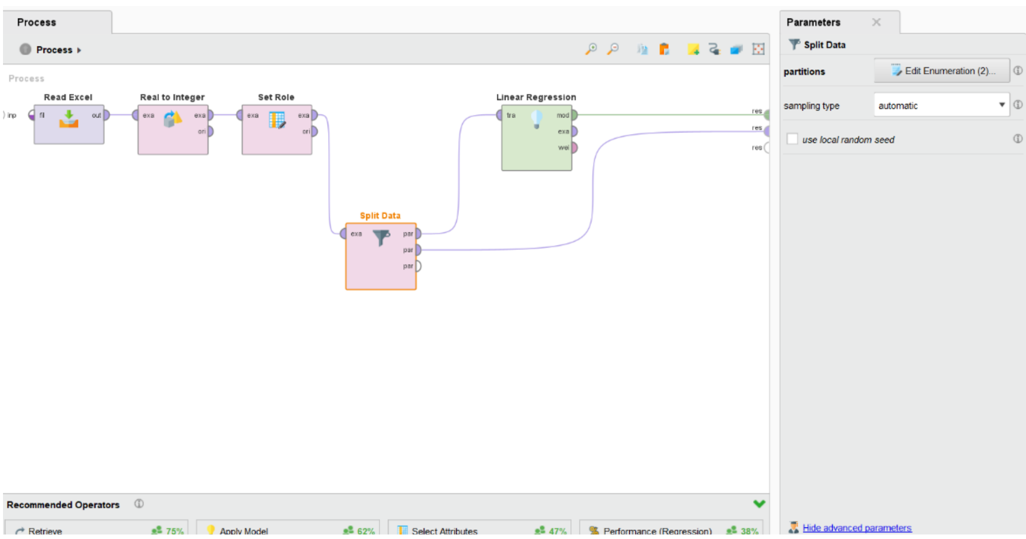


Figure 24: Linear regression operator

4. Now using the apply model operator connect the split data to the other split data. Run the model and then view the results in the result view.

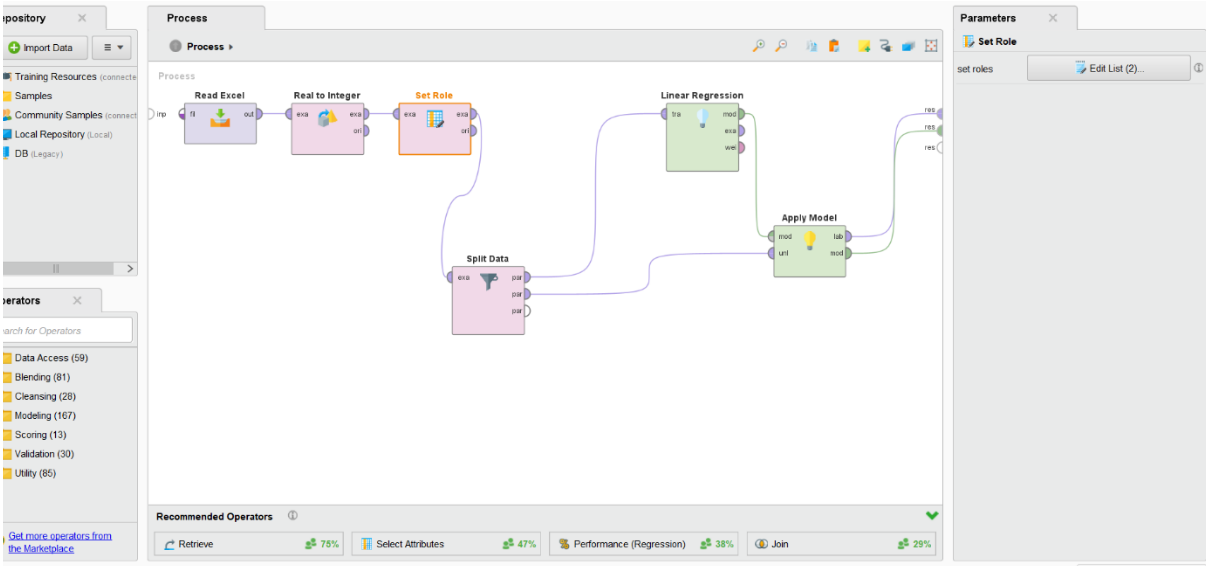


Figure 25: Apply model

Evaluation

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Agriculture, forestry...	-0.607	0.194	-0.017	0.402	-3.130	0.002	***
Mining and quarrying	1.545	0.217	0.102	0.056	7.117	0.000	****
Manufacturing	1.972	0.262	0.215	0.017	7.520	0.000	****
Electricity, gas and ...	5.617	0.398	0.169	0.092	14.131	0	****
Construction	1.988	0.351	0.061	0.106	5.667	0.000	****
Finance, real estate...	2.235	0.193	0.480	0.006	11.580	0	****
(Intercept)	-4853.117	6308.785	?	?	-0.769	0.444	

Figure 26: Linear regression output

Linear regression is about predicting how close an observation is to an imaginary line representing the average, or centre of all points in the data set. This imaginary line gives the first part of the term linear regression To calculate the prediction with linear regression is y=mx+b. The y variable is the target variable which is the GDP at market prices. The x variable is the given predictor attribute or also known as the independent variable. Since the focus is on the manufacturing industry performs, this will be the x variable. The m variable will be the coefficient shown in the second column which is 1. 972.The coefficient is the amount of weight the attribute is given. The manufacturing sector has the 4th heaviest attribute showing it has a significant impact on the GDP market prices. The b variable is constant and will be the intercept, which is -4853,177. (North, Linear Regression, 2018)

The question is how much of a contributor is the manufacturing industry to the GDP at market prices. The market price is the 4th biggest contributor to the manufacturing industry. The biggest being electricity, gas, and water, then fiancé and then Construction. (I feel like you can add more but this answers this question

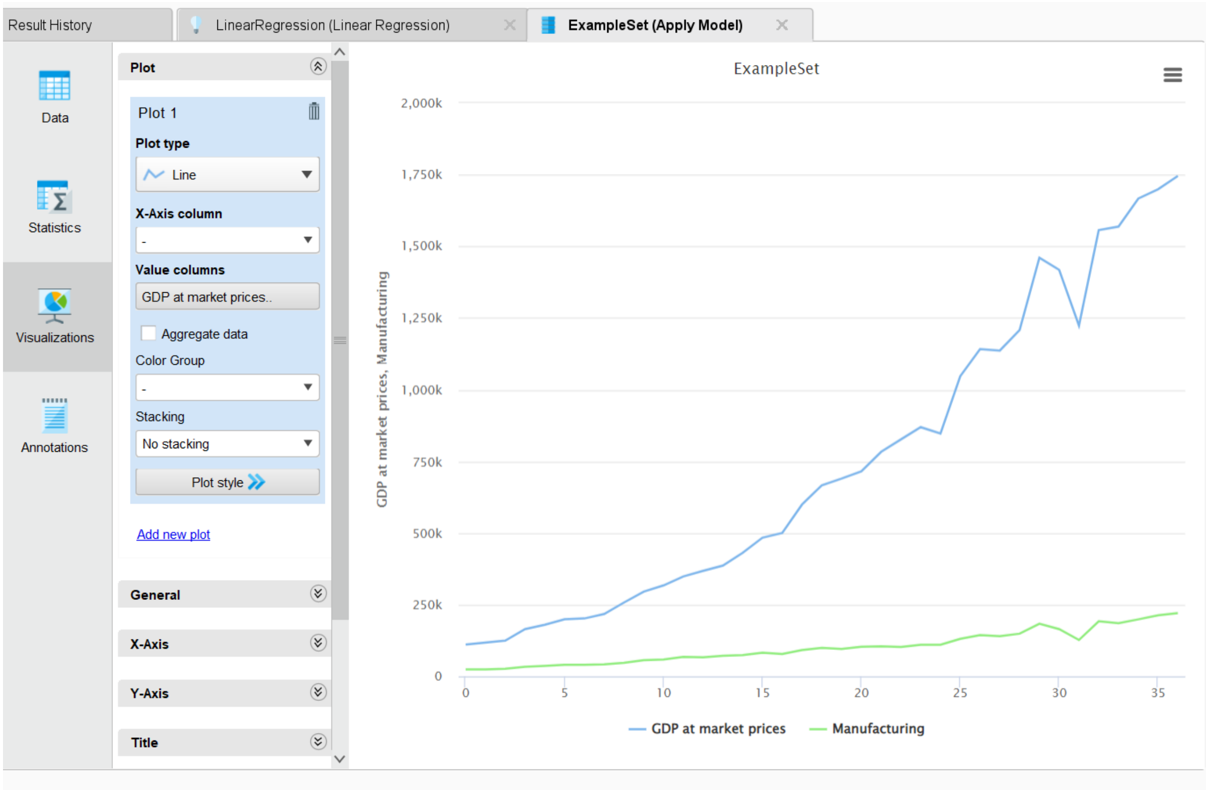


Figure 27: Linear regression apply model visualisations.

Although the variable that needs to be considered is the manufacturing industry, using the GDP as a predictor is important because when the GDP goes up so does the manufacturing industry this can be seen by the apply model graph.

Deployment

The manufacturing sector is the 4th biggest contributor to the GDP. This means that that the manufacturing sector is a valuable industry to invest in, in South Africa. Ongoing monitoring and evaluation of the deployed linear regression model's performance needs to be ensured. Updating the data as new statistics are released needs to be conducted to improve prediction accuracy overtime. Additional features that may enhance the model's ability to make predictions needs to be explored. It also necessary for investors to continuously monitor the GDP to see how it performances as the GDP says a lot about how the manufacturing industry is performing.

Conclusion

The Crisp-DM cycle that has been deployed is used to understand the feasibility and viability of investing in the manufacturing sector for robotic household appliances in South Africa. The data gathered and analysed is used to investigate employment statistics, manufacturing sales and GDP contributions. The objective guided the analysis by only focusing and assessing the manufacturing landscape and potential investment opportunities in South Africa. To make actionable insights for an informed decision-making process it was necessary to investigate employment rate, household appliance sales and manufacturing's impact on the GDP.

The data used came from three different datasets related to employment, GDP and manufacturing sales which was sourced from Department of Statics South Africa. Through these dataset insights about the manufacturing sectors performance and its relevance to the proposed investment was gained.

The first cycle which is a classification model is a correlation matrix. Correlation analysis allowed for meaningful relationships between manufacturing employment across provinces and potential investment opportunities. Gauteng and Kwa-Zulu Natal showed the highest positive correlation which indicates that it has the best conditions for manufacturing investments. Correlation analysis highlighted which provinces has the highest potential for investment in manufacturing. It is necessary to have ongoing monitoring and evaluation as well as data updates to ensure long-term success.

The second cycles deployed is predictive models. The predictive models are gradient boosting and Linear Regression. Gradient boosting model predicted future trends in household appliances manufacturing sales by using historical data and performance metrics. The model has a decent performance. This is shown by the low error metrics and a reasonable R-squared value which gives insights for production planning and resource allocation.

The linear regression model analysed the significance of the manufacturing sector in its contribution to the GDP at market prices. The relationship between manufacturing industry performance and GDP showed that it has the 4^{th} largest coefficient indicating that it is the 4 largest contributing industry to the GDP. This allows investors to see that the manufacturing sector is a thriving sector.

Reflection

The datasets are all linked to the manufacturing the household robotic appliances but not supplying them in South Affrica. The data that was collected for supplying went against investing for supplying the product. The different models used gave insights into the manufacturing landscape in South Africa that provides investors with decision-making insights. Data about technology adoption was either not available without a pay wall or was non-existent. There also wasn't data about consumer behaviour. This investment in manufacturing the products the best root. The manufacturing sector in South Africa is a big contributor in south Africa, so exploring the datasets gave insights into why investors should invest in the manufacturing sector.

Investors, South Africa, and its people will benefit from this. Investors will get labour at a good price due to the low exchange rate and South Africa's manufacturing industry is the fourth greatest contributor to the GDP , thus by the business investing to manufacture their products in South Africa will in turn grow the GDP. Also, there will be an increase in the employment in the manufacturing sector. The manufacturing industry does have a tendency of going down and up, this needs to be considered and by such household appliance sales can also go down. The models need to continuously update to ensure real-time analysis and accurate analysis.

References

Geeks for Geeks. (2024). *Gradient Boosting in ML*. Retrieved from https://geeksforgeeks.org/ml-gradient-boosting/

KZN Top Business Porfolio. (2024). *Manufacuring Secotr*. Retrieved from KZN Top Business Porfolio: https://kzntopbusiness.co.za/site/manufacturing

North, M. (2018). Correlation Methods. In M. North, *Data Mining for the Masses* (pp. 71-79).

North, M. (2018). Linear Regression. In M. North, *Data Mining for the masses* (pp. 161-168).