**Business Analytics 2: Decision Making with Data 2024 V2**

**Assignment 3 (Data Analysis- Part 2)**

**Answer 1:** Alteryx software helps with predictive analytics in a variety of ways. It offers end-to-end processes, from data cleansing and manipulation to model development, analysis, and performance validation utilising a variety of performance measures.

The first step in data cleaning/manipulation is determining the data type required for predictive analytics. The "Select" tool allows you to edit the metadata of the data and set the metadata based on the data. For example, a numeric field with whole integers has its metadata set to 'Double'. Using the "Select" tool, you may change the metadata to 'Int64', ideal for running predictive analytics on this data.

The next step in the data cleaning/manipulation process is to identify the data, evaluate it, and verify for discrepancies. In this case, if there are any null rows or columns, use the "Data cleaning tool" to replace them with blanks for V\_strings data types or zeros for numeric data. Alteryx has the advantage of allowing you to clean or edit data even further using tools like as formulas, filters, sorting, and so on.

Once cleaned and altered, the dataset is ready for analysis. The analyst will use their models to the data to perform predictive analytics. Analysts can analyse their data using a variety of predictive technologies. Analysis tools include linear regression, logistic regression, decision trees, and forest models. Once the analyst has run his predictive tools, he may evaluate and confirm their effectiveness using model metrics such as R-squared value for regression models, which indicates how well the data fits into the regression model. For example, suppose the regression model has an R-squared value of 0.71. It signifies that 71% of the data fits well into the model and is very robust. This can also be visualised in Alteryx using regression model charts. Aside from R-squared, you can validate performance using MAE (mean absolute error) for regression models. The same can be done for classification models using the Confusion matrix and F1 scores.

The Alteryx platform has extensive predictive analytics features, allowing users to traverse the process more efficiently. However, one noticeable issue is the platform's handling of huge data files, which can cause substantial slowdowns. This issue frequently causes analysts to endure delays or completely restart the program, resulting in losing momentum. Furthermore, Alteryx lacks built-in visualisation tools, requiring manual data export to external visualisation platforms for complete analysis. Moreover, the tool's cost may be a barrier for small and medium-sized businesses (SMEs), as license prices vary and may not be within their budget.

Answer 2:

1) Linear regression is used for regression tasks. It is a supervised learning algorithm. This method compares the target (dependent variable) to its predictors (independent variables). For example, linear regression can predict a car's average kilometre mileage per litre depending on size and weight.

2) Logistic regression is a supervised learning algorithm. It is used for binary classification problems with only two possible outcomes: yes or no, true or false, 0 or 1, etc. Using logistic regression, the bank can determine the borrower’s borrowing capability by analysing earnings, liabilities, credit score, and other factors. Hence, banks could use this to approve or reject loan applications.

3) Decision Trees are a supervised learning algorithm. This predictive tool is primarily used for classification tasks, using the algorithm to predict an outcome. An e-commerce company can use this tool to predict whether a customer will buy based on the customer’s browsing behaviour, purchase history, etc.

4)Support Vector Machine (SVM) is a supervised learning tool for classification purposes. This is an excellent tool for banks to segment their customers into categories. They can use the SVM tool to classify Loan applicants into different segments, such as high-risk and low-risk customers, by analysing data for income, marital status, employment, profession, credit score, and other metrics.

5) K-means is an unsupervised learning tool that groups data into K clusters based on similarity. For business analysis, K-means can be used to categorize or subcategorize customers based on their purchase history, money spent, and frequency of purchase, such as grouping them as "High Value," "Mid-Range," and "Seldom Buyers."

After assessing the data and aligning it with the business goals of predicting automobile sales, I chose linear regression to run my analysis. A linear regression model can offer precise, valuable insights into the variables influencing sales when used to forecast car sales, which is why I chose it for my analysis. I ran an association analysis tool to determine the correlation between sales and confirm whether I had chosen a suitable model. In this process, I decided on “Sales” as a target field and chose different variables to find any correlation.

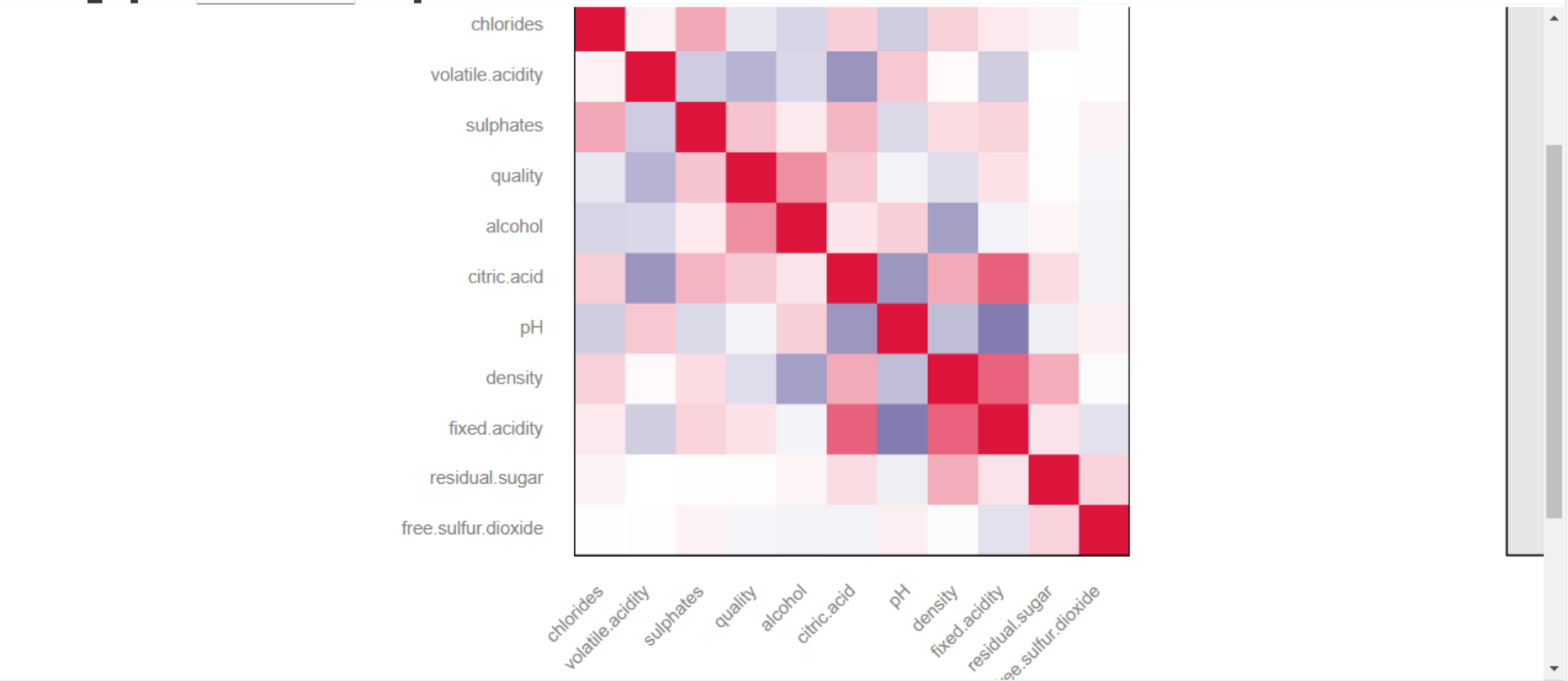
After reviewing the data, my first task was to clean and reorganise the metadata which is part of data cleaning/manipulation. I utilised the "Select" tool to modify the metadata of all columns based on their value, i.e. changing numeric data set up as V\_strings data into numerical data for the study. During this process, I identified an issue with the 'order date' column metadata, which I resolved by converting the order date column. The metadata of the order date column was originally set up as V-string data, which I changed to date using the Date-time tool. In the process of data cleaning/ data manipulation, I noticed that there are additional independent variables, which are V\_string data types like deal size and product line, but can be used for analysis; however, they must be converted into numeric values. To address this issue, I converted the 'deal size' and 'product line' columns into numeric values using the Python tool, aiming to enhance the accuracy of the linear regression analysis. After converting the 'deal size' and 'product line' columns into numeric values, I used the "Join" tool to combine both columns with the original dataset before performing the linear regression analysis. After completing the relevant steps above, I ensured the goal of the business, understood the data, and prepared the data by cleaning and manipulating it in all instances required so that I was ready to use the Linear regression tool to run my analysis.

My next step was to begin the process of model development. I dragged and dropped the linear regression tool into my workflow and started the model development phase. In this phase, I chose “Sales” as my target variable and “Quantity Ordered,” “Price Each,” “Order Line Number,” “Days Since Last Order,” “Product Line,” and “Dealsize” as my predictors. The relationship between predictors and targets can help to justify sales with the help of predictors. This allows businesses to create new strategies, make informed decisions and forecast accurately.

**Answer 3**: (Provide screenshots of data, model, explain selection of variables, and interpretation of the results):

**Question 3:**

For the wine quality dataset, I first assessed the initial goal of predicting wine quality based on the variables in which I reviewed each column to see which could be my predictors. After that, I used the “Select” tool for data manipulation & cleaning by changing the metadata of the columns to ‘double’ for numeric columns with decimal places for more accuracy. I also changed column names for ‘Quality’, which was not in numeric value, to quality classification for classification analysis. After this process, I used the association analysis tool to predict my variables further. I used Quality (numeric) as my target variable and selected all the other variables for correlation to see which predictors I could use for regression and classification analysis. As per Pearson product-moment correlation, all the chosen variables had some level of correlation with wine quality.



I used association analysis and logical reasoning and used all the remaining features as my independent variables to run my analysis. Using wine quality as my target variable, I ran linear regression, a decision tree, SVM and a neural network for my analysis.

**Decision tree**

* Accuracy – 86.6% of data was predicted accurately. Which is a good prediction.
* F1 score –53.3% means that this model was not the best at predicting wine quality.
* Precision – 72% of data was predicted correctly, which is a good prediction for this model.
* Recall – 42.4%, which is relatively low, means the model did not find actual positive cases accurately.A screenshot of a computer

  Description automatically generated

**Neural Network**

The neural network is overall a good fit, as the graphs below show. The graph proves that most of the predictions made by this model are close to zero, which means they are close to their actual values, with very few predictions to the left or right of the model.

A graph of a graph

Description automatically generated with medium confidence

Using this neural network, it is evident that the model is suitable for predicting wine quality. The confusion matrix image below indicates that the model made incorrect predictions for high-quality wine 18 times. It was accurate in predicting medium quality 514 times. In the same category, it also predicted low quality 22 times and high quality 44 times as medium quality, which are minor mispredictions.

A screenshot of a computer

Description automatically generated

The Effect plots below suggest the relationship between the quality of the wine (high). As per the images below, the fixed acidity plays a part in predicting high wine quality. The lower the fixed acidity, the higher the wine quality and likewise for volatile acidity.

A screenshot of a graph

Description automatically generated

A comparison of a graph

Description automatically generated

A graph of a function

Description automatically generated with medium confidence

**SVM**

The SVM model is equally good at predicting wine quality compared to neural networks. Using SVM, it is evident that the model is suitable for predicting wine quality. The confusion matrix image below indicates the model's incorrect predictions for high-quality wine 8 times. It was accurate in predicting medium quality 524 times. In the same category, low quality was predicted 18 times, and high quality was predicted 37 times as medium quality, which are minor mispredictions.

A screenshot of a computer

Description automatically generated

**Linear regression**

As per the image below. Here is how this model performed.

* R-Squared: 0.419 - This means that 41.9% of the data fits well into this model, which is considered low. - The
* adjusted R-squared is 0.409, equating to 40.9%, indicating that some predictors do not contribute significantly to the analysis. - The
* Mean Absolute Error (MAE) is 0.477, suggesting that the model's prediction was off by 0.477 units from the real values. A lower MAE indicates a better model. - The
* Mean Absolute Percent Error (MAPE) for this model is 0.088, or 8.8%, which is considered low and indicates that the model’s prediction is off by 8.8%. -
* Mean Squared error is 0.385, providing information about how far off our predictions are. -
* The Root Mean Squared error is 0.621, indicating that the average error in our predictions is 0.621 units. -
* F-Statistics—41.23 with 11 and 628 degrees of freedom suggests that the model is good at prediction, as the value is high. -
* RSE is at 0.626 on 629 degrees of freedom, meaning, on average, the model's predictions deviate by 0.626 units from the real values.

A screenshot of a computer

Description automatically generated

According to the linear regression interpretation, the volatile acidity P-value is less than 0.001, indicating high significance. The estimated value is -0.986346, signifying a negative impact. In simpler terms, as volatile acidity increases, the quality of wine decreases, showing a negative association, which is also the same as Chlorides.

The presence of sulphates and alcohol is positively associated with the quality of the wine. The p-value for sulphates and alcohol is highly significant, as indicated by (\*\*\*). This indicates that higher levels of sulphates and alcohol lead to increased wine quality.

Total Sulfur dioxide and pH hold some significance due to their p-value being only one start (\*) and having a negative association. This means that an increase in total sulfur dioxide and pH will decrease the quality of the wine.

**A screenshot of a computer

Description automatically generated**

**Question 3 workflow below**

**A screenshot of a computer

Description automatically generated**

**Answer 4**: (Provide screenshots of the performance of all models and explain, and interpret them):

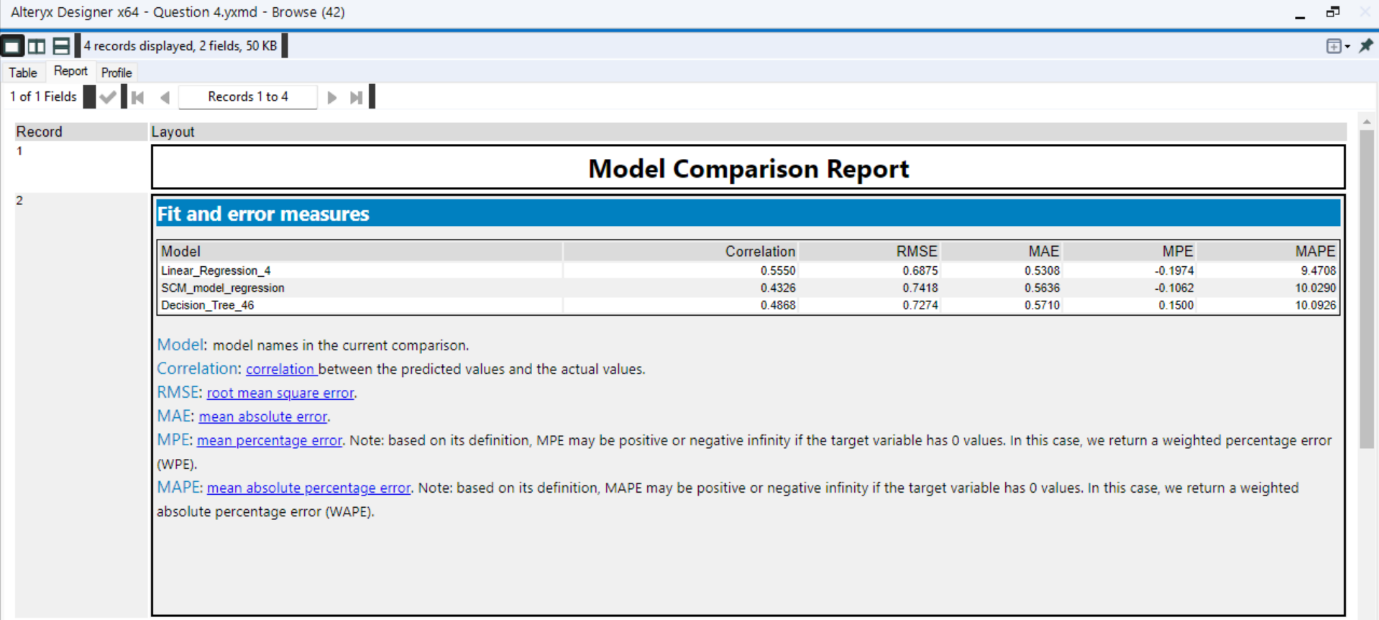
I conducted two model comparisons to obtain the best results for the analysis and evaluate the accuracy of the models. One comparison involved regression analysis, where I considered linear, SVM, and Decision Tree models. The other comparison was for classification analysis, where I opted for Decision Trees, SVM, and neural network models. When running the classification analysis, I changed the target variable from the numeric 'Quality' column to a string variable 'Quality classification'.

The SVM model performed the best for classification analysis, with the highest overall accuracy compared to Decision Trees and neural networks. It also outperformed in predicting high categories, as indicated by the Accuracy\_High score of 0.3580. Additionally, the F1 score for SVM was better than the other models, demonstrating a better balance between precision and recall across all classes.

A screenshot of a computer

Description automatically generated

According to the data in the image, the linear model demonstrates the strongest correlation between the predicted and actual values for the regression analyses. The linear model also achieved the lowest RMSE at 0.6875, indicating that it has the smallest average errors when compared to other models. Additionally, the linear model has the lowest MAE at 0.5308, implying that it makes the smallest average error. This is further supported by the MAPE in the image, which is smaller than other models at 9.4708, indicating a small percentage error.



The image above proves that Linear regression is the best-fit model for running regression analysis.

I will document every stage of the model-building process, which includes data preparation, model selection, and evaluation. I plan to share the code along with detailed comments and the dataset on a platform like Microsoft SharePoint for collaboration. Additionally, I will create a detailed manual or tutorial to guide others on how to use and modify the model. I recommend experimenting with more complex models, refining existing ones, improving data quality, and considering other variables to improve forecast accuracy and overall performance. I

**Workflow for Q4**

**A screenshot of a computer

Description automatically generated**