Cindy Hernandez
CIS 3120
Professor Jairam
10 May 2024

Viewing the Dairy Farms

The dairy industry is one of the leading industries that produces the most greenhouse gas emissions in the world. The performance of the dairy industry is closely linked to various factors, with price and quantity being amongst the most significant. For this project, I wanted to use a linear regression to see the relationship between the price per unit and how it affects other factors. Dairy is a food item that is constantly being purchased, whether it be milk, cheese or yogurt. By seeing these relationships, it would help those within the industry glimpse at how the market works and what improvements can be made.

For creating the linear regression model, I started with importing the necessary libraries to build our model and evaluate the data points. I used panda, numpy, scikit-learn, matplotlib and seaborn. Matplotlib and seaborn were used to visualize the data, sckit-learn was used for the machine learning, numpy for our operations and pandas to store and handle the data. I created a data frame that would load the data from the csv file and display the rows and columns of data.

Data cleaning and preprocessing are simple yet important steps that involve dropping irrelevant columns or missing data rows that would change the outcome of our analysis. For this analysis, I dropped the columns 'Total Land Area (acres)', 'Location', 'Number of Cows', and others which were descriptive, but not needed for the analysis we are conducting. I also looked for null or missing values within the dataset, however, the dataset I worked with had none so I was able to continue with my analysis. I then printed the descriptive statistics which would provide insights into the data, such as the mean, standard deviation, and quartiles. These statistics are crucial for understanding the data distribution and the central tendencies of the features. These statistics aid in understanding the data's behavior and aid in the following steps.

The next step I did was defining the feature set (known as X) and the target variable (known as y). For my analysis, I wanted to use 'Price per Unit' as my target variable and have my predictors be the quantity, total value (refers to the total amount of dairy), the quantity sold, price per unit (sold), and the reorder quantity. This separation of points is crucial for the training and testing portion of the model, since X would be our predictor variable, and y is the variable we aim to predict. To split the data into the training and test sets, I used a 70-30 ratio and shuffled the variables to allow for reproducibility. By using this split, we can simulate the model's performance within real-world application.

The linear regression model is trained by instantiating it and fitting it to the training data. This step allows the features and the target variable to be related to one another. When the training process is complete, the test set can be used for making predictions based on which its effectiveness in performance and behavior can be measured using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics, as well as many others. These estimates give insight into how well a given model has been able to predict future outcomes. The MSE was a 7.78 and RMSE was a 2.78. I also requested a detailed summary that includes the coefficients, R-squared value, and p-values. These are important for viewing and understanding the significance of each predictor variable as well as the overall fit of the model. This detailed analysis aids in interpreting the model's effectiveness and the importance of the features to predict the target variable. In the summary, the overall r-squared value was a .988 meaning that our analysis was a good fit and has a strong correlation.

The Mean Squared Error (MSE) measures the average of the squares of the errors, meaning that within our model it calculates the average squared difference between the actual and predicted values. In this project, the MSE value was calculated to be 7.78. A lower MSE value usually signifies a better fit, as it means the predicted values are close to the actual values. However, it may also incorrectly classify and calculate larger errors, which can cause issues within the outliers.

The Root Mean Squared Error (RMSE) is the square root of the MSE and provides an error metric in the same units as the target variable. In this analysis, the RMSE was 2.79. RMSE is useful because it provides a measure on how well the model's predictions match the actual values. RMSE is easier to understand and interpret than MSE because it provides an absolute measure of fit. A lower RMSE indicates a better fit, with the value suggesting a moderate average prediction error for our variable 'Price per Unit'.

Visualization plays a crucial role in interpreting the model's performance and locating potential issues. The plots I created to analyze the regression was a scatter plot of actual vs. predicted values to visualize the accuracy of the prediction, residuals plot to check the distribution of errors, a histogram to display the frequency distribution of the variable, and a heatmap to visualize the correlation among the features I tested. These visualizations are crucial for seeing issues with our features that are non-linear and understanding our data and model performance.

One of the visualizations I used was a scatter plot of the actual versus predicted values for 'Price per Unit'. This plot helped assess the accuracy of the model's predictions. The scatter plot revealed that most of the points were clustered around the line of perfect fit, meaning that it has a strong correlation and is a good match. This also indicates that the model's predictions closely match the actual values. Overall, this visualization great for understand the relationship between the two features and how they are represented in this analysis.

Another essential visualization is the residuals plot, which displays the residuals, meaning the difference between actual and predicted values, against the predicted values. The residuals plot in this project showed a random distribution of residuals around the horizontal axis, indicating that the model's errors are randomly distributed. This randomness suggests that the model does not suffer from systematic errors and that it captures the underlying data pattern well. It also demonstrates a constant variance, meaning that will not increase or decrease over time since the points stay relatively in the middle.

The heatmap of the correlation matrix provides insights into the relationships between different features in the dataset. In this case, the heatmap revealed that some features had a strong correlation with 'Price per Unit', while others were less significant. For example, the feature 'Total Value' displays a moderate to high correlation with 'Price per Unit', demonstrating that as the total value increases, the price per unit tends to increase as well. On the other hand, ‘Reorder Quantity’ showed no correlation to the target value, as well as zero correlation to other features such as quantity, total value, quantity sold, and price per unit sold. All in all, the correlation matrix displayed good result between the points, but also visualized the issues between some of the features as well.

Lastly, the histogram of 'Price per Unit' displayed the frequency distribution of the target variable. This histogram showed that 'Price per Unit' had a normal distribution, with most of the values staying within the same range. Since it is a normal distribution, it means that the data points stay frequent near the mean and tend to decrease as they move farther away. Since most of

these data points stay around the same area, I believe there is a low probability that it will skew left or right in the future.

One of the challenges I faced in this regression was choosing the features I wanted to use for my regression. I had already set my mind to using the price per unit and chose the features with values to analyze. However, I also wanted to test the farm size and cows to see if those features would affect how much a product would be priced. I also wanted to test the approximate total revenue; however, I wasn't sure how relevant it was to my analysis and how I could create and visualize the relationship between the two.

In conclusion, building a linear regression model involves a step-by-step process, from preparing the data, and then training and evaluating the model. Each step is crucial for ensuring model accuracy and reliability. Without these features, our model would not be able to make clear analyses that many people can use. By following these steps and using the proper techniques, a well predicted model can be created to make accurate predictions for any problem. This process highlights the importance of data cleaning, proper model evaluation and testing, as well as visualization in within linear regression. The end product would ultimately lead to a model that can display valuable information and predictions based on the data.