Data 301 Project

Analyzing an online retail store's transaction history

Yohan Sofian

ysofian@calpoly.edu
Student
Department of Engineering
California Polytechnic State University
San Luis Obispo, CA

Connor Gamba

cgamba@calpoly.edu
Student
Department of Statistics
California Polytechnic State University
San Luis Obispo, CA

Abstract

The following research report represents analysis conducted on an online UK-based retail store, evaluating its transaction history and logs, to hypothesize store performance and extract data on the store's customer base. Our analysis for this project was conducted and processed using jupyter notebooks in Google Collab, allowing for feature engineering of variables and models to be produced and easily replicated. Our interest for this project stems from our curiosity to see whether the sales data for a store could provide meaningful insight into future sales growth, or could provide recommendations for potential action that the store could take to optimize sales and revenue streams. Additionally, we were curious if analysis of the sales log could provide valuable insight into the store's customers, which if so, we believed could allow the store to have a better understanding of how it can address and cater to its customers.

INTRODUCTION

For our project, we were interested in studying transaction logs for an online store as we thought it would be interesting to see if we can accurately make predictions on store performance or draw meaningful conclusions about the store's customer base. Such conclusions would provide insight into potential ways the store could optimize its revenue, such as by targeting customers that are more likely to be frequent buyers or by learning things such as how certain products perform and what can be done to better the performance of such products. Thus for our project, we collected an extensive transaction log that contained information about every given sale or return for the store. In doing our analysis, we ultimately found that despite the limited information provided with the data set, through extensive feature engineering, our team was able to answer many of our questions and gain valuable information on the store's customer base, and performance throughout the year.

DATA

The data set that we used was the *Online Retail II* data set sourced from and available on the UCI Machine Learning Repository. This dataset is from an all-occasion giftware wholesaler/store in the United Kingdom. It contains information of numerous transactions to other businesses in the world from 2009 to 2011. Due to the large scale of the data, we had difficulty working with it especially for machine learning since we do not have the computational power to process all of the data quickly.

Unfortunately, even though there is a lot of data in the dataset, there is only information about invoice number, stock code, product name, quantity ordered, invoice date, customer id, and country where the customer is from. As a result, we did a lot of feature engineering such as transforming data from the invoice to separate day, month, and year, grouping some data together to compute analysis per month, quarter, and year.

QUESTIONS STUDIED

We group questions that we plan to study together, which includes basic analytical questions about the data, simple regression models, clustering, and classification.

Questions that involve simple analytical analysis are as follows:

- What item is the best seller in every month based on the dataset?
- What item is the best seller in every quarter based on the dataset?
- Which region is this business most successful in? Calculate the total sales for each country and find the biggest sales.
- Which region is this business least successful in? Calculate the total sales for each country and find the lowest sales.
- What is the largest transaction on an invoice based on the dataset?
- What is the mean revenue of this business per month?
- Calculate total revenue for this business for each month based on the dataset
- Calculate the average number of transactions per month for this business.
- Calculate the average total sales on the invoice for this business.
- Predict the average revenue per month by multiplying the average number of transactions (invoices) with the average total sales per invoice.
- What is the best selling product for each country?
- What is the product that generates the most revenue for each country?

Questions about simple regression models:

- Build a linear regression model that best predicts the total sales per month for this business by using appropriate variables. Compare it to the previous prediction. Calculate the MSE and R2 Score.
- Use K-means to cluster the data and show a scatter plot of the data with a plot of price vs qty.
- Build a KNN regression model that best predicts the total sales per month for this business by using appropriate variables. Use hyperparameter tuning to determine the best value of K to use in the model. Compare it to the previous predictions and results from the linear regression model

Questions about clustering:

- Can customers be segmented based on purchasing behavior (frequent buyer, seasonal buyer, occasional buyer, etc.) through the use of clustering?
- Show a plot of between quantity a customer purchased for a given month and total amount spent for that month
- Limit plot size to remove extreme observations and allow for break-down of clusters
- Plot quantity a customer purchased for a given year and total amount spent for that year
- Comparing Quantity Per Month to Quantity Per Year
- Comparing Quantity Per Month to Quantity Per Year excluding unusual observations
- Comparing Amount Purchased Per Month to Amount Purchased Per Year
- Comparing Amount Purchased Per Month to Amount Purchased Per Year excluding unusual observations

Question about classification:

- Can we determine if a customer will be a repeat customer based on purchase history using classification models? **(**Use the oldest 80% of the sales data in training a classification model, and use the most recent 20% as a validation set for evaluating accuracy of our predictions. Use feature engineering for whether or not there was a repeat customer.)

ANALYSIS AND EVALUATION

Basic Analysis:

A lot of the basic analysis questions involve asking simple things about how the business is performing. Most of it is done with pandas and numpy, without any complicated models, regression, or anything like that. This involves questions such as how much revenue this business generates per month, per quarter, or per year; what region this business generates most/least revenue, what is the largest transaction per month, and other basic questions like these. We intend to get to know how the business is performing financially and what product is the best seller for this particular business.

Regression Models:

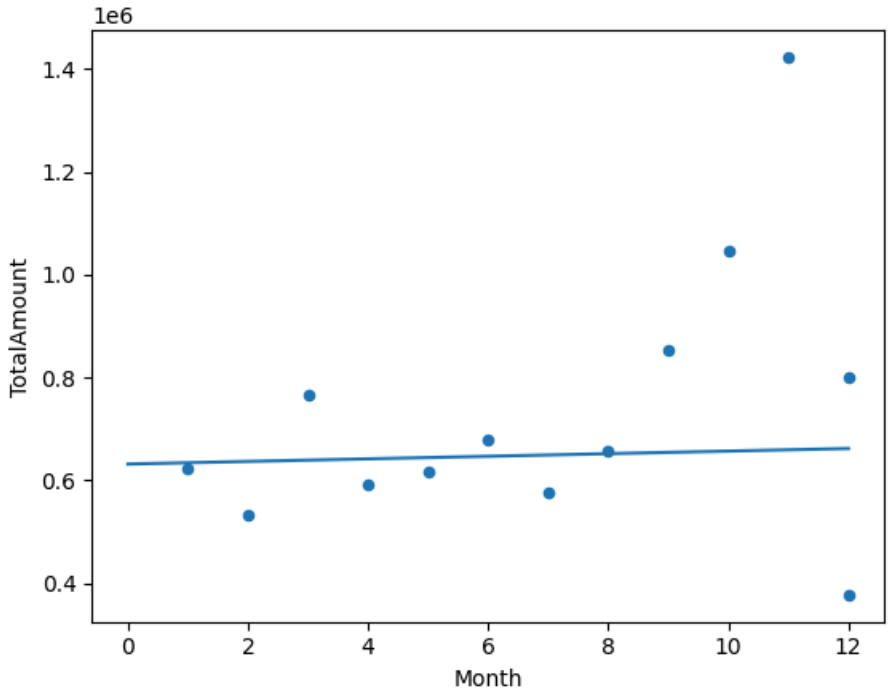
We have a total of 3 regression models. The first model is a part of the basic analysis, the other two use different machine learning models. Each model is explained below.

Basic Linear Regression:

- The purpose of this basic linear regression is to predict how much revenue this business generates per month. This linear regression is done by calculating the average total number of invoices/transactions that this business has and the average amount of each invoice. For example, if the business has 50 invoices per month and for each invoice the amount of transaction is say 1000 GBP, the total revenue of the business per month is 50.000 GBP.

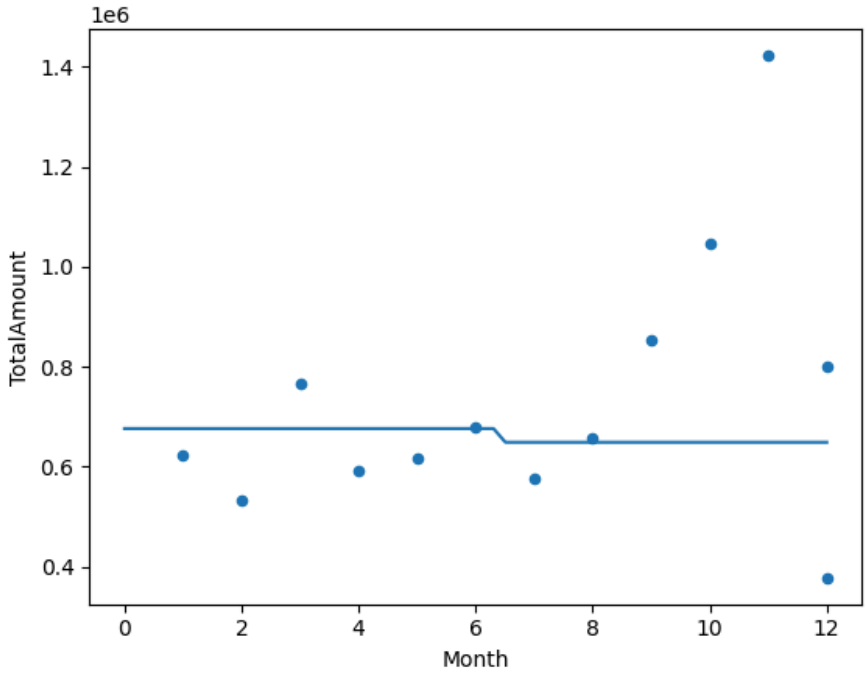
Linear Regression using Sklearn:

- Our goal is to create a better linear regression model to predict the revenue for the business per month by using multiple variables to predict and using the Linear Regression library from Sklearn. We were using only month as a predictor so it turns out to not be accurate. The MSE score is enormously large. The model depicts a relatively flat line, thus it seems this model could be useful in estimating the average monthly revenue for the overall year, but fails to be valuable in prediction of revenue for a given month.



Regression using K-means:

- We then tried to use K-means for regression and unfortunately still unable to really predict accurately the revenue that this business generates per month. We were also doing hyperparameter tuning, but we feel that the data does not really fit linearly even with multiple variables involved. The below model depicts our implementation after hyperparameter tuning, and as can be seen the result is very similar to the linear regression. The graph depicts that this model may hold some value in determining the average monthly revenue throughout the year, but fails to be valuable in actually predicting revenue from month to month.

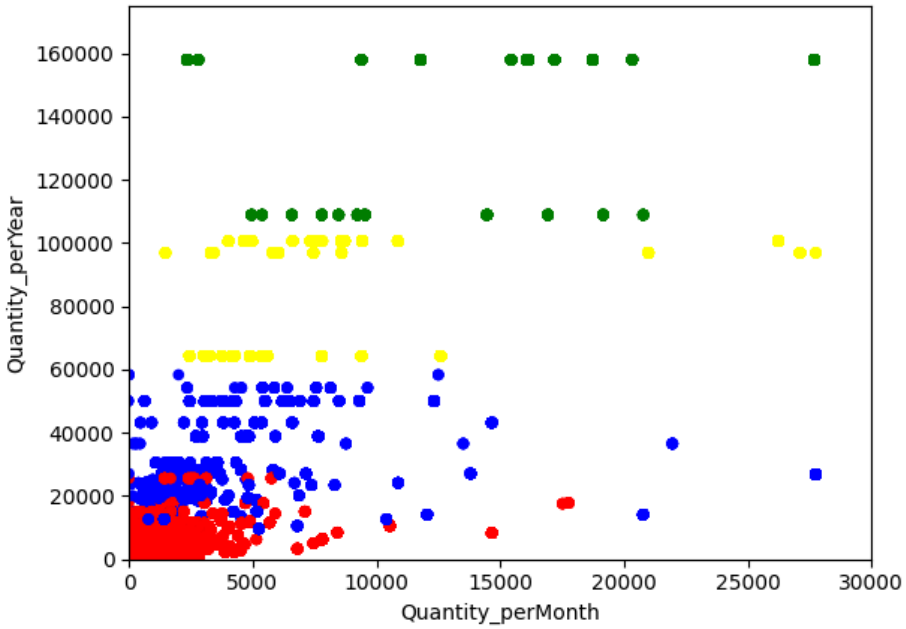


Overall, our regression models were not able to accurately predict sales for a given month, though they did definitely seem to succeed in predicting what average monthly sales were for a given year, as the trend lines for both KNN and Linear Regression are overall pretty consistent to our original average monthly sales prediction/calculation. Explanations for our model's performance likely stem from our lack of historical data, being that this data set only contains sales data for three years, thus it is hard to make accurate predictions as we are greatly confined by our data points. Another challenge with this model is that we are trying to predict monthly sales for a given month, yet our data set has 3 months and corresponding sales for every month, as the set contains 3 years of data. This poses a challenge as essentially, for every x-value (month) our model has 3 different total amounts to work with. This is why in the production of our models we averaged the sales data for each month, so that one month will provide the average sales data of that month from the past three years. Thus by default, the scatter plot does depict a fairly decent prediction as points depict average monthly sales data from the last 3 years. However, our intent with the regression models was to accurately produce a trend line for anytime throughout the year, which could allow for analysis of total revenue for half of a month for instance.

Clustering:

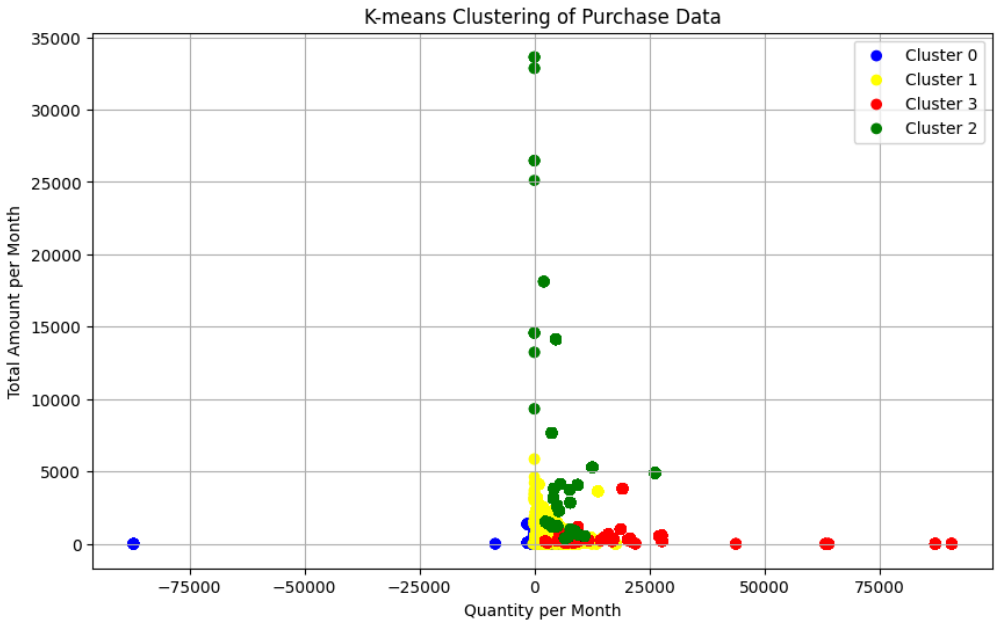
Our research involved using two different clustering models to better understand the store's customers and sales trends

The first clustering model we implemented was intended to see if we can categorize customers based on the type of buyer they may happen to be such as if they are an occasional or frequent buyer. In order to evaluate this, the variables **Quantity_PerMonth, TotalAmount_PerMonth, Quantity_PerYear,** and **TotalAmount_PerYear** were feature engineered to specify how many items a customer buys per month and per year, in addition to total spend. Using these categories allows us to determine if a customer buys infrequently, such as if they often make few purchases on a monthly basis but make large purchases overall in the span of a year.
To implement a k-means model for this data set, we used 4-clusters as looking at scatter plots depicting the relationship between **Quantity_PerMonth** to **Quantity_PerYear** this amount of clusters seemed to provide us with the most reliable and distinct groupings. Viewing various scatterplots composed of the 4 variables we used in our model, we can see that cluster 0 (blue) tends to be composed of low-moderate quantity per month buyers, with moderate quantity per year, relatively low yearly spending with low to moderate monthly spending. Thus this cluster seems to be composed of more frequent buyers, who overall are not large buyers relative to the other groupings, though do make regular purchases. Cluster 1 (yellow) depicts a customer who tends to purchase a relatively high quantity per year, with relatively low quantity per month. Spending for this customer seems somewhat proportional in terms of monthly spending and yearly spending. Thus this customer is likely more infrequent with the actual items they purchase, likely buying less expensive goods in higher quantities in times when the quantity is high. Cluster 2 (green) depicts



very high quantity per year and high spend per year, with relatively low quantity per month and moderate spend per month. This is likely an infrequent or occasional buyer, who makes many purchases and more costly ones on an occasional or rare basis. Cluster 3 (red) shows a proportionate trend in comparing the month variables to the year ones, and overall seems to be composed of people who in general do not buy many items from the shop.

Our second implementation of a clustering model using this data set aimed to provide insight into item similarities. In other words, we wanted to see if clustering could be used to perhaps categorize or pair items that are frequently sold together. Such insight could be useful for a store to have as it could provide them knowledge as to how and what items can be bundled to boost revenue and sales. This was implemented by using our same base predictors from the previous clustering model, and again clustering based on these same variables, with each point now depicting a product from our store. By comparing quantity per month and total amount spent per month for each particular item, we are now able to evaluate items based on how they are purchased and in what quantity to get a sense of item similarity. From our model, Cluster 0 (blue) depicts return items, as the quantity is negative for all items in this cluster in our scatterplot, thus, this cluster depicts items which are perhaps frequently returned. Cluster 1 (yellow) depicts items which are sold in relatively low quantities per month and generate somewhat moderate sale for the company. Thus, these appear to be mid tier or mid priced items. Cluster 3 (red) depicts items sold in a high monthly quantity with a low overall total amount or sales revenue. Thus, these are likely low value items that are sold in high numbers. Finally, cluster 2 (green) depicts items sold in low monthly quantities with relatively high to high total amounts per month. Therefore, these items appear to be high value or more expensive items, which are not sold in high numbers, likely due to the cost factor. These are probably also more occasional type products being that they are high cost and sell a low quantity .



Classification Model:

Our classification model's goal is simple, which is to predict if customers will be a repeat customer based on the previous purchases. This turns out to be a pretty accurate predictor since it guesses correctly most of the times if someone will be a repeat customer.

This dataset unfortunately has missing values, and some of the invoices have negative items (which means it was returned). Thus, we had to delete and adjust some of these missing values and returned items invoice accordingly.

At first, when we are training the data, it seems like it takes forever. We ended up having to use only 1% of the data for training, and it turns out to be sufficient enough to predict whether a customer will be a repeat customer or not. Using 5% of the data takes longer than 30 minutes to train. We then scale the data, and using the K-neighbors classifier (n=5), we fit the data, and get 97% accuracy.

CONCLUSION

In conclusion, there are a lot of things that we can still learn from this dataset

despite the lack of other important information for a business. We were able to look at the best-selling products for each month and quarter, which region the business did good and bad, which can help this business itself to decide on keeping inventory, or make decisions to expand marketing to do better in other parts of the world.

We also tried building a regression model with three different ways, and ultimately we believe that it is pretty hard to predict a revenue based only from the month; maybe experimenting with other variables such as demand for a certain product, a holiday season, and so on might generate a better prediction. This however will make this analysis way too complicated and might include things that we have not learned from this class. We also perform clustering to simulate how different customers are divided, and so on.

Our classification problem has also successfully predicted if a customer will be returning to buy other things from us with 97% accuracy. It is also pretty safe to say that running a business as a wholesaler will probably have a lot of repeat customers.

REFERENCES

1. Chen,Daqing. (2019). Online Retail II. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D.