Term Project Report on Restaurant Chain Data

David D Berberena

Bellevue University

DSC 550 Data Mining

Brett Werner

June 1, 2024

Term Project Report on Restaurant Chain Data

**Introduction**

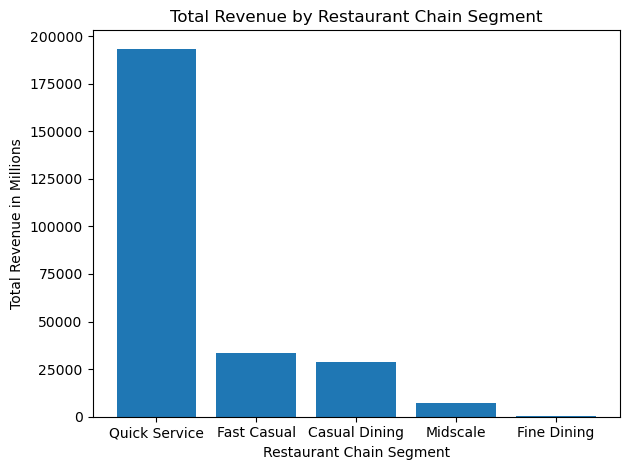
Investing in the restaurant industry here in the United States can be very lucrative for the American investor. With the abundance of restaurant chains offering a variety of options to consumers, there is sure to be a return on investment hidden within the number of establishments serving countless amounts of sustenance. However, if an asset management company specializing in the management of customer funds allocated towards stocks, mutual funds, options, ETFs, and other marketable securities in the foodservice industry markets themselves as being able to provide a stable and/or high rate of return, the company must know how best to manage their customers' assets to provide the most convenient type of ROI possible for the client. For this to be realized, the company must understand where the restaurant industry's money is made so they can invest their customers' funds to make money for both their clients and the company itself.

This is where restaurant chains come into play, as these giants dominate the revenue generated by the foodservice industry. The issue of confidently identifying lucrative restaurant chains to invest in, whether individually or through mutual funds and ETFs that encompass many of these top restaurant chains as part of a restaurant sector, can be solved by exploring data related to these chains, their revenue, and their sector designation to understand the rate of growth. If asset management companies geared towards foodservice industry asset investment can successfully understand and analyze related data to make an informed decision for their clients, the company itself will ultimately increase its bottom line and retain clients who will continue to realize the ROI that they are interested in. As stakeholders in the finance industry, it would be in their best interest that the companies that they are invested in have access to data and solutions derived from the data and the ability to analyze and sift through data to ultimately prosper in the future with data becoming increasingly more important within all facets of business.

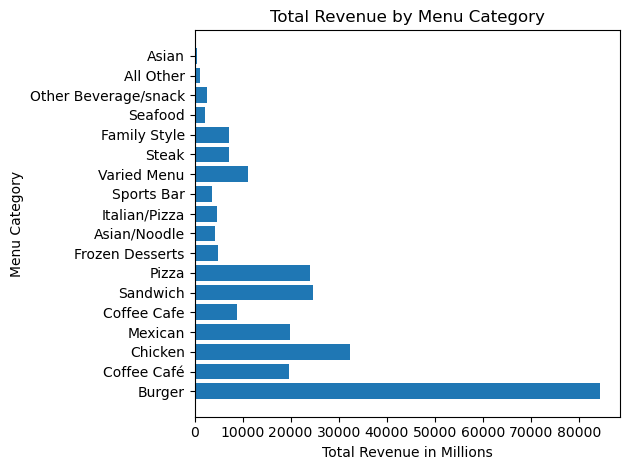
With regards to the data I wish to peruse concerning restaurant chains in the foodservice industry, I have retrieved data from Kaggle outlining information concerning the top one hundred restaurant chains in the United States in 2021 based on 2020 performance (Harshith, 2021). Using the data, I will be analyzing what sector of the restaurant chain subcategory of the foodservice industry an asset management company should invest their client's money into to realize a strong ROI. To help address the solution to this topic, a few questions can be asked of the data. What restaurant chain sector generates the most revenue? Which sector has the highest positive change in sales? What about the highest positive change in units? How about inquiring even about the relationship between menu category and total sales? I also will be looking to see if there is a strong linear relationship between a change in the number of restaurant chain locations and a change in total sales. Questions that can be asked relating to this data inquiry are whether there is a correlation between these two variables, how strong that correlation is if one exists, and how the total sales are affected in the future provided an increase or decrease of chain locations. To explore the potential answers to these scenarios, we can delve into the data that I have found.

**Exploratory Data Analysis**

Looking at the dataset Kaggle provided, I have managed to put together a couple of key visualizations to begin answering some of the questions above (Harshith, 2021). To craft charts using the data, I simultaneously needed to transform the data and plot it. Some transformations that needed to be performed were the removal of special characters within the observations and the coercion of the observations to their correct data type. The most crucial visualizations that I have crafted directly answer the question of which type of restaurant chain is the most lucrative to potential investors. The below images are these visualizations, one focusing on the most lucrative restaurant sector, and the other highlighting the most lucrative restaurant menu category.



Looking at the results of this bar chart, it is abundantly clear that the quick service sector of the top 100 restaurant chains generates the most revenue in comparison to the other sectors. With total sales measuring almost eight times more than the next most lucrative restaurant sector, it would be a wise choice for the asset management company to encourage clients to invest in the quick service sector should they desire to partake in being a shareholder within the foodservice industry. The next visualization will handle answering the question of which restaurant menu category brings in the most revenue among the top 100 restaurant chains in the United States in the year 2020.



With the output of the horizontal bar chart, the menu category that brings in the most revenue is the Burger category. After reviewing the above visual aids and the other visualizations reported in the initial data analysis section within Jupyter Notebook while considering the initial business problem at hand, there are many conclusions that can be drawn from the data. If the asset management company is working with a client who wishes to achieve maximum short-term ROI, the clear answer for the company would be to invest the client's funds into securities dealing with quick service burger restaurant chains. These chains have the highest revenue, least negative percent change in sales, and is the median sector in terms of percent change in units. If the company is dealing with a client who is looking more stable long-terms investments, then that client's assets would be best placed into those securities focusing on fast casual burger chains. These chains have the highest percent change in units, the second highest sector in revenue, and the median sector when looking at change in sales. Each client's needs and wants are different, and to stay relevant and prosper in the data-oriented world we live in today, companies need to be able to analyze the data to find an answer for each customer they do business with for maximum retention and long-term financial growth and stability.

Now that the visualization stage of my restaurant chain data dive is complete, I needed to transform the data further to prepare for the model generation stage. These transformations included further removal of special characters that prevented the correct data types from being output, further coercion of the incorrect data types to have the observations be of the correct type, and feature reduction so that the dataset only contains the most relevant features to answer the questions regarding the relationship between the change in number of brick-and-mortar restaurant chain locations and total sales. With the data now ready for the model generation stage, the model choice that seemed most appropriate to answer the question of linear correlation between these two variables was a linear regression model. Within the data, there was another variable that described the change in restaurant chain location. In addition to the actual number of locations that differed from the previous year, I wished to use the variable that contained the percentage of change between restaurant chain units from the previous year. With these three variables now in play for the model to work with, the linear regression model would now be a multiple linear regression model as two variables would function as predictors while the total sales variable would be the outcome variable.

One of the assumptions the data and this model makes is that there is a linear relationship between the predictor and outcome variables. One can easily assume that the more restaurant locations a chain has, the higher the total sales. The data also suggests that the more positive the change in units and sales, the higher the total sales. Looking at the correlation matrix I created in Jupyter Notebook, I saw that there is a strong positive correlation between the number of restaurant chain units and the total sales. This indicates a strong linear relationship, so the linear regression model is a sound model choice. The change in units and change in sales variables show a weak positive correlation, so while the relationship isn’t exaggeratedly positive, it still fits within the linear regression model’s assumptions. With the data primed for model generation, I had the data split into training and test sets to fit the linear regression model to see whether the model’s performance confirmed the earlier assumptions.

**Conclusion**

Interestingly enough, the multiple linear regression model that was created was not a good fit for the data. This was made apparent by its abysmal performance metric values, which were the root mean squared error (RMSE), mean absolute error (MAE), and the R-Squared (R2) values. With the R-Squared value clocking in at 0.4, almost sixty percent of the variance in the data is left unexplained by the model. Both the RMSE and MAE values were very high, signifying that the error between actual and predicted observations was rather high. As these metrics confirm major issues with the model’s performance with the data, the model was not ready for deployment. My recommendation to anyone planning to continue with this model would be for them to look for more data while testing to see whether the hyperparameters of the linear regression model need to be tuned to reduce the chance that the model is overfitting the data. Obtaining more data on restaurant chains not in the top 100 in the United States allows the model to have more data to train the model and test it. Looking at the hyperparameters allows for confirmation of the linear regression model as being the best model choice for the obtained data.

A challenge that I foresee coming up however, is the availability of more relevant data and whether one can find it quickly. As the data being worked with represents restaurant chains in 2020, it may prove difficult to find credible data that aligns with the timeframe of the current data. However, the inclusion of new data can also be considered an opportunity depending on how the original hypothesis is altered. If more data is found regarding these restaurant chains a couple of years before or after the current data, we could then be working on a much different, yet just as intriguing model. Restaurant chain data from two separate years would help the model predict future observations with a much higher accuracy than using one year’s worth of data alone. Overall, I believe that this project has shown that even if data looks to be meaningful and related, making sure that enough data is present is key in confirming how the data looks along with its assumptions. While there is no specified size data for a project such as this should be, it is apparent that the model’s performance metrics spill its secrets of whether the data is adequate or not, prompting further data mining in the future.

References

*Harshith, J. (2021, November 14). Top 100 biggest restaurant chains 2021. Kaggle. https://www.kaggle.com/datasets/johnharshith/top-100-biggest-restaurant-chains-2021*