Term Project Milestone 3: Preliminary Analysis

David Culhane

DSC 630-T301

Andrew Hua

***Will I be able to answer the questions I want to answer with the data I have?***

The data to be used for this analysis will most definitely accomplish the goal of making sales predictions. The main question will be if the models employed are able to predict the weekly sales total of a given department of a given store. The data to be used in the models have data regarding store, department, the week, sales data, and other data that can affect sales. These fields should be able to work together to predict sales numbers.

***What visualizations are especially useful for explaining my data?***

Line charts and stacked area charts would likely be best for explaining aspects of the data. Line charts will be able to all the various aspects of sales and aspects that affect sales on a two-dimensional plane. Any plot could tie a couple factors together and a third could be employed by use of color. Stacked area charts would be able to see store-wide sales by stacking the department sales numbers of each week on top of each other. No matter the graph, the aspect of time must be preserved if possible.

For any line chart though, it would be advised to use no more than a couple of departments in any plot. The sheer number of store and department combinations would make any plot attempting to plot them all entirely unreadable. Ideally, only one graph per pair of store number and department would be used and individually analyzed.

I used a scatterplot to analyze how the different store types experience different amounts of sales. No differentiation was given for individual departments and it was clear that each store type had clear areas, allowing for the decision to be made to split the data by the type of store to make sure that the models used are not compromised by the store type.

***Do I need to adjust the data and/or driving questions?***

The data has been slightly adjusted. All three dataset components have been merged to make sure each line in the resulting set has a sales number for a department as well as store and department number and influencing factors like temperature, gas prices, and unemployment. The three different store types had different sales profiles, so they were each given their own dataframe to be used for modeling.

The fields related to anonymized sales data unfortunately had to be removed. All five of those fields only had data for dates after November 2011. That meant each field was missing more than 60% of its data. My fear was that imputing these fields would dramatically skew the model if those fields were incorporated. To avert that, those five fields were removed.

***Do I need to adjust my model/evaluation choices?***

I will not need to adjust my model choice since the objective has not changed even though we are moving from a possibility of one model to a certainty of three models to use. A linear regression or LOESS model would likely be the best model to use.

The main question will be at what date in the data is the line drawn to split the dataset into training and testing sets. Random selection is not an advisable path since factors in the data are time-dependent. It could make sense to use the 2010 and 2011 data as part of the training set and 2012 data as the testing set even though the 2012 data ends in October.

Model evaluation should proceed as planned, though use of a ROC chart and its AUC will not fit the objective of this analysis and prediction since they lend themselves to categorical target variables. R2 and RMSE scores will still provide assessments for the predictions made by the models.

***Are my original expectations still reasonable?***

I believe that my original expectations are reasonable. The store and department fields will be turned into dummy variables for each dataframe before being given to the models. The factors affecting sales at each store over a given week (temperature, gas price, CPI, unemployment, and time of year) will be used from the testing data to help predict the sales after the model has been trained. These predictions can be scored to evaluate the model’s performance and judgements can be made from those scores.***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***Milestone 2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Predictive analytics can be applied throughout various business spaces in order to give a business an idea of what to expect in the future. All a business needs to apply predictive analytics is vast amounts of data, either acquired through its own records organically or by purchasing it as a business asset. Retail environments specifically can make use of predictive analytics to inform almost every aspect of their business operations to forecast when customers will be shopping, what these customers will be buying, how much staff should be on hand, when to order more of any individual product for their inventory, and the list could continue to go on. During the period I spent working in retail, I was able to access the result of predictive analytics in real time, our store’s sales goals. These predictions would be the basis for staffing decisions and be a measuring stick for how our store performed for management.

If a major retail chain is able to use predictive analytics to predict the performances of its stores, that gives it an edge when it comes to making predictions for how the company should report to its shareholders and make a litany of other related decisions. For this reason, I have chosen sales prediction/forecasting as the topic for my term project. The data I plan on using for making these predictions comes from Kaggle at <https://www.kaggle.com/datasets/manjeetsingh/retaildataset>. The dataset itself spans three Excel CSVs, a features dataset, a sales dataset, and a stores dataset. Together, they describe weekly sales data across 45 retail stores across different regions from 2010-2013. Each store has its own departments and sales information.

The features dataset contains fields for the store number (anonymized to numbers 1-45), the date the week starts with, the average temperature in the store’s region for the week, the cost of fuel in the store’s region that week, five markdown fields used to mark sales taking place after November 2011, the value of the consumer price index in the store’s region that week, the unemployment rate in the store’s region that week, and whether or not the week is considered a holiday week. Across these fields there are 8,190 rows of data. The stores dataset contains information about the type and size of each store. Lastly the sales dataset has fields for store number, department number, the date the week started with, the department’s sales that week, and whether the week was considered a holiday week. Since there are 45 stores each with their own departments, the sales dataset is much larger than the features dataset, clocking in at 421,570 rows of data to use.

With this glut of data, we will need to learn a model or set of models through either store-wide sales approach or department by department approach. Since the nature of our target variable, sales, is continuous, a linear regression model will likely be the approach taken. LOESS modeling and linear modeling with polynomial functions could also come into play since they can follow the nature of sales is not necessarily linear. Additional exploration in this space can be explored. The goal will be to see if our model(s) can successfully predict the sales figures towards the end of the dataset. Since the data we are using is chronological, I intend to split the data into training and testing sets by year, with the last year of data serving as the testing set and the rest of the data serving as the training set. If the data is split randomly as most other situations would have me do, there is a chance that some time periods through the year may not be captured sufficiently by the model(s).

I do recognize, however, that there is a risk of overfitting for the model with a couple years of complete data being used in the training set. To test the model, predictions can be made using data from the features dataset regarding dates, fuel, temperature, and other non-sales numbers. Given the continuous nature of the target variable, scoring metrics of choice should include R2, root mean squared error, and ROC curve and its area under the ROC curve.

The data being used does not contain customer information, information that could hint towards what retail chain this data is about, or anything else of the sort. Therefore, there should not be any ethical implications with this endeavor in predictive analytics.

If this data doesn’t end up working out for this analysis, I could pivot from analyzing data about multiple stores to data from a single store with data about product IDs and other anonymized information, found on Kaggle here: <https://www.kaggle.com/datasets/abdullah0a/retail-sales-data-with-seasonal-trends-and-marketing>. This dataset has around 30K rows and would provide a more granular look into the same phenomenon.

Hopefully the size of this dataset doesn’t prove to be an issue. Most of the data is numerical and with only some categorical data. Some of the categorical data has numerical values, so those will need to be re-categorized as such during data preparation. Afterwards, creating the final dataset will require a couple merges to make sure only one dataset needs to be created. My hope is that merging onto the sales dataset will be the correct course of action. That way, the final dataset will have store, department, weather and other non-sales data, and the sales-oriented data in a single set to be used for model creation. It would also cap the size of the dataset at the size of the sales dataset, 421,570 rows. The last thing needed would be even more than 421K rows of data to feed into any given model.

One concern with these datasets is that they do not all span the same timeframes. The features dataset spans from 2010-2013 while the sales dataset spans from 2010-2012. The solution to this problem is likely to be to only model and predict using data from 2010-2012, with 2012 data being the testing data.

# References

M. Abdullah, S. Y. (2024). *Retail Sales Data with Seasonal Trends & Marketing*. Retrieved from Kaggle: https://www.kaggle.com/datasets/abdullah0a/retail-sales-data-with-seasonal-trends-and-marketing

Singh, M. (2017). *Retail Data Analytics*. Retrieved from Kaggle: https://www.kaggle.com/datasets/manjeetsingh/retaildataset