MGT 6203 Group Project Proposal Template

**TEAM INFORMATION (1 point)**

**Team #:** 1

**Team Members:**

1. Yizhi Zhang; harryzyz

I am a Data Scientist living in Chicago, IL. I earned my PhD degree in Applied Mathematics at Illinois Institute of Technology. My job focuses on Credit Risk Model and Business Sending Model.

1. Dawn Shi; dawndriver

I am living in Austin Texas. I terminated my career after working as quality engineer for a giant oil and gas company for 9 years due to family relocation. I graduated from National University of Singapore, major in mechanical engineering, and moved and stayed in United States since 2016.

1. Christian King; 42ck2009

I am a process engineer living outside of Birmingham, AL. I have a Bachelor of Mechanical Engineering from Auburn University. I have been working in automotive manufacturing for the last five years. I do a lot of data visualization for manufacturing data.

1. Madison Smith; smithmn918

I live near Birmingham, Alabama with my husband and 2 young children. My undergraduate degree is in Statistics from Brigham Young University. I currently work as a Model Validation Analyst at a large regional bank, where I focus on fraud and consumer lending models.

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:** Retail Store Sales Forecast and Accuracy Analysis

**Background Information on chosen project topic:**

Sales forecasting is one of the most critical elements of business planning, and companies invest considerable funds to get more precise prediction, as inaccurate prediction could lead to loss of customer, loss of revenue, or even fail of the business. Recently the retail giant Target is facing dilemma of profit hit and stock be downgraded, one of the reasons is because inaccurate forecasting causing overstocking and raising of operation cost(Market Watch, 2022).

Accurately predicting future sales is perhaps the most difficult part of the revenue forecasting process. Accurate forecasting could also improve the inventory management and customer experience, thus, to maintain or increase the market share. Companies utilize different models to predict future sales, plan for stocking, and perform logistic analysis. The most common models used on the market are Gut-Feel Forecasting, Almanac Forecasting Method, Funnel Forecasting, Portfolio Forecasting, Multivariate Regression Analysis, and Machine Learning/AI. Companies select the model that best fits the business depending on market scale and forecast budget. Among all these methods, Multivariate Regression Analysis and Machine Learning/AI are the two methods that are adopted by most of the big companies. Machine Learning/AI is the most sophisticated method, and not all companies can afford it, on the other hand Multivariate Regression Analysis is comparable, but more cost-efficient and affordable (People AI, 2022).

There are three common forecast traps we could fall into, one of them is lack of datapoints (Score, 2022). Therefore, Walmart past sales is chosen for our study, this world top retail company will have sufficient datapoints for team to study with.

Another common trap is the Over-Simplification Trap. Confident entrepreneurs typically project growth linearly with consistent growth over time. However, this creates a ‘hockey stick’ graph that the forecast will have stellar growth that keeps going up and up indefinitely (Score, 2022). Thus, the project will approach by plotting and identifying dominant variables and creating dummy variable and interaction terms, if necessary, to avoid model over-simplified.

The last trap is the Not-Based-in-Reality Trap. To ensure fairness of the model (Score, 2022), data from different states and stores will be evaluated together to avoid cherry-picking data.

When talking about the datasets, to find a relatively accurate regression model that can predict the future store sales, a few factors need to be considered:

* Item Price changes due to promotions and/or events (Score, 2022)
* State factors due to different incomes and preferences
* Store location can make sales vary
* Etc.…

A few data cleaning and wrangling methods need be clarified, like how to cluster data points and categorize different events, grouping by states and/or by stores. Past research (Seaman, 2022; People AI, 2022) will be studied and referenced along with materials from the course of MGT 6203.

To summarize, this project is to study accuracy using linear regression to forecast future sales of the retail giant Walmart, by combining historical data gained from Kaggle, and technologies obtained from the course of MGT 6203. The result will be evaluated based on a different dataset obtained from the same Kaggle competition.

**Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):**

This project is to explore the possibility to use linear regression method together with cluster approach (Seaman, 2022) for sales forecasting, by using regression techniques and statistic knowledge gained from course of MGT 6203. A dataset about Walmart past sales acquired from past Kaggle competition will be used for modeling and analysis. The model obtained will be evaluated with another dataset containing out-of-time sample Walmart sales. In the meantime, explanatory variables will be compared by their influence on the sales. Please refer the following sections for more details.

**State your Primary Research Question (RQ):**

Can we accurately forecast sales in terms of revenue for the retail giant Walmart based on week number, special events, store ID, and/or item price using linear regression techniques covered in MGT 6203?

**Add some possible Supporting Research Questions (2-4 RQs that support problem statement):**

1. Which explanatory variables are most important in forecasting sales in terms of revenue?
2. Can we simplify the forecasting process by aggregating the data at a certain level, while still forecasting sales revenue accurately?

**Business Justification:** **(Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)**

As stated in the background information, sales forecasting is one of the most critical elements of business planning and companies invest considerable funds to receive more precise predictions. Inaccurate sales predictions could lead to loss of customer, loss of revenue, or even failure of the business.

Accurately predicting future sales is perhaps the most difficult part of the revenue forecasting process. Accurate forecasting could also improve inventory management and customer experience, thus, to maintain or increase market share.

Further, there are three common forecasting traps that we must avoid:

1. Lack of data points
2. Over-simplification
3. Not-based-in-reality

This problem is interesting to solve because the business implications are real and substantial. Additionally, there are several pitfalls to avoid in the forecasting process.

**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

The data can be found at the following link: <https://www.kaggle.com/competitions/m5-forecasting-accuracy/data>

**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

The data consists of hierarchical sales data from Walmart stores located in three US States (California, Texas, and Wisconsin), as displayed in Figure 1. Sales data is provided at the item, department, category, and store level. The data includes the following key information:

* Time series data
  + Daily unit sales of 3,049 items classified into 7 product departments (Foods 1 – 3, Hobbies 1 – 2, and Household 1 – 2) and 3 categories (Foods, Hobbies, and Household)
* Explanatory variables
  + Sales price
  + Promotions
  + Day of the week
  + Month of the year
  + Year
  + Week number
  + Special event name
  + Special event type
  + State ID
  + Store ID
  + Department
  + Item ID

The data was pulled from 1/29/2011 to 5/22/2016 and daily sales are forecasted for the following 28 days – 5/23/2016 to 6/19/2016. We plan to use a few of the explanatory variables outline above to estimate sales in terms of revenue for retail goods sold in the United States between 5/23/2016 and 6/19/2016. Our group will not utilize the time series data to ensure that we can appropriately use the tools covered in MGT 6203, specifically linear regression.

Further, rather than generate sales revenue predictions for each product at each store, our group plans to aggregate sales by week and possibly an additional hierarchical variable – store, department, or product category. Our group will work to determine the aggregation level (or lack thereof) during the exploratory data analysis (EDA). At the store level we would generate weekly sales revenue for 10 individual stores for 4 weeks (28 days). At the department level, we would generate weekly sales revenue for 3 departments within 10 individual stores for 4 weeks. At the category level, we would generate weekly sales revenue for 7 product categories within 10 individual stores for 4 weeks. We will likely aggregate the data to: (1) simplify the problem and (2) practice data wrangling skills. Without aggregation, we would generate weekly sales revenue for 3,049 products within 10 individual stores for 4 weeks (28 days).

Figure 1: Hierarchical Data

A picture containing chart

Description automatically generated

The datasets include:

1. calendar.csv: Contains information surrounding the dates on which products were sold, as displayed in Figure 2.
2. sales\_train\_validation.csv: Contains historical daily unit sales by product and store, as displayed in Figure 3.
3. sell\_prices.csv: Contains information about the sale price of a product by store and date, as displayed in Figure 4. The sales price does not change within the week, rather week to week.
4. sales\_train\_evaluation.csv: Contains historical daily unit sales by product and store, including 28 days for forecasting. This data is identical in content to the data displayed in Figure 3, with 28 additional days.

Figure 2: Calendar Data

Table

Description automatically generated

Figure 3: Historical Daily Unit Sales Data

Table

Description automatically generated with medium confidence

Figure 4: Sale Price Data

Table

Description automatically generated

**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

As stated above, the data includes the following key information:

* Time series data
  + Daily unit sales of 3,049 items classified into 7 product departments (Foods 1 – 3, Hobbies 1 – 2, and Household 1 – 2) and 3 categories (Foods, Hobbies, and Household)
* Explanatory variables
  + Sales price
  + Promotions
  + Day of the week
  + Month of the year
  + Week number
  + Year
  + Special event name
  + Special event type
  + State ID
  + Store ID
  + Department
  + Item ID

The explanatory variables outlined above will be considered independent variables, while weekly sales revenue will be considered the dependent variable. This problem also contains a time series aspect. Time series data is a collection of observations that are indexed in time order. The primary goal of time series analysis is to predict future values of the time series variable. Therefore, the independent variable in a time series analysis is generally time. As stated previously, our group will not utilize the time series data because time series tools will not be covered in MGT 6203. To align the project more fully with materials discussed in the course, our group will focus on predicting weekly sales revenue based on the explanatory variables, excluding time.

Our group will transform the event variables into 3 indicator variables based on event type (sporting, cultural, national, and religious). We will also need to determine the best course of action when multiple events occur in the same week.

Our group hypothesizes that the most important explanatory variables will include:

* Week number
* Special events
* Store ID

Our group hypothesizes that price will also be an important explanatory variable if we do not aggregate the data above the product level.

**APPROACH/METHODOLOGY (8 points)**

**Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))**

Firstly, we will define the problem and reformat the data to suit the problem definition. We will explore data with visualizations to achieve some basic sense of the data. We will combine the information from the price dataset and the sales dataset to compute the daily revenue at each store. Then we will extract weekly revenue data at each store from the daily revenue data.

Secondly, we will create features according to the problem. We will consider special events data from the calendar dataset and pair the events data with the revenue data. We will clean the data with small tricks to help get a model with good performance and generalizability.

Thirdly, after the data preparation, we will start making predictions in intuitive ways. For example, we may take the mean of previous sales as the prediction of the future sale. Afterwards, we will proceed to build an analytical model using linear regression.

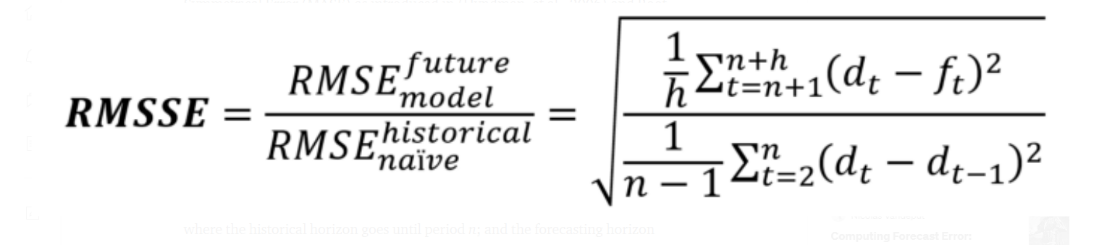
After the model construction, we will define the evaluation metrics and assess the model. To fine-tune the model, we will use cross-validation to optimize the data starting date, eliminate insignificant features, and rebuild the model.

To compare multiple models against each other, the first step will be to plot the residuals and compare the adjusted r-squared values. If further comparison is needed between models, we will use the error measured against the true values from the evaluation data. This could be done using RMSE, MAE, or average relative error.

**Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement**

The expected outcome of the analysis is to predict sales revenue over the 28-day period within the specified category. At this time, it could be by individual store, every department in each store, or every product category in each store. We would expect to produce an estimate for sales revenue in each category (store, department, or product category) that is close to the actual sales revenue for that time frame within a reasonable confidence interval.

We will use RMSSE to give a measure of fit for the model onto the last four weeks (28 days) of the evaluation data. We will try to obtain the lowest RMSSE possible. The RMSSE is chosen because it is scale independent. It can be used to measure across series with different scales. For example, the model may predict one category extremely well, but that category may only have 5 sales per week, while the model may be inaccurate in a category that has 10,000 sales per week.



We could also use the true value compared against predicted values to measure our results.

Some examples may include: RMSE, maximum error, or average error.

Good modeling will lead to accurate predictions that can forecast future sales based on current data. We will consider a relative error within the range of 20% being successful.

**What business decisions will be impacted by the results of your analysis? What could be some benefits?**

Accurate predictions of future sales will greatly help retail companies to making future promotion plans, adjusting logistic arrangement, and more importantly, generate more profit in business. Thus, this analysis is very beneficial for business owners.

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

The project will take about six weeks, from Jun/6/22 to Jul/20/22. We divide the project into three parts. From Jun 8 to Jun 22, the team will review the raw data, define the problems to be solved, and make plans about how to clean the raw data. From Jun 23 to Jul 6, the team will perform preprocessing, build the initial model, and prepare for the progress report. From Jul 7 to Jun 20, the team will tune the model, optimize hyper-parameters, make the predictions to our problem, and write/present our final report.

**References**

*Market Watch*. (2022, 6 18). Retrieved from https://www.marketwatch.com/story/target-stock-downgraded-as-multiple-analyst-groups-blame-execution-for-profit-hit-11652985474?mod=mw\_quote\_news\_seemore&mod=article\_inline&mod=article\_inline

*People AI*. (2022, 6 18). Retrieved from https://people.ai/blog/sales-forecast/#:~:text=Multivariate%20regression%20is%20a%20statistical%20method%20of%20sales,have%20a%20corresponding%20effect%20on%20the%20predicted%20output.

*Score*. (2022, 6 18). Retrieved from https://www.score.org/resource/how-poor-forecasting-can-sabotage-your-business-plan

Seaman, B. (2022, 06 19). Retrieved from https://forecasters.org/wp-content/uploads/gravity\_forms/7-c6dd08fee7f0065037affb5b74fec20a/2017/08/Seaman\_ISF-2017.pdf