Assignment 3

Jiaxing Lin

Metropolitan College, Boston University

AD571 A1: Business Analytics Foundations

Prof. Hyunuk Kim

Assignment Due Date: 06/19/2023

Prior to conducting analysis on the dataset, I excluded data that could distort the results, specifically data with sale prices or square footage equal to zero, ensuring a more statistically accurate representation. By extracting the year of sales and aligning it with the relevant building code IDs, I refined the data set to encompass only residential property types. I then focused the scope of my analysis on data from 2009 onwards for the assigned neighborhood HARLEM-CENTRAL.

Part 1

To understand the complex relationships between variables and provide a more comprehensive picture of the data dynamics, I applied predictive analytics with a multiple linear regression model. Initially, I transformed the building code ID column into factor variables or categorical variables so it can be further included in the model. Subsequently, I developed a multiple linear regression model to examine how the independent variables of the property's construction year, building code ID, gross square footage, and the number of residential units could be utilized to explain the sale price of the residential properties. Summarizing the results provides detailed information on the model, including the coefficients that indicate the expected impact on the sale price for each unit change in the variables.

Furthermore, I can determine the most and least valuable predictors in determining the sales amount. Based on the p-value, the year of build is the least valuable predictor. The p-value is way outside the 95% confidence interval and statistically insignificant, meaning there is not enough evidence proving that there is a relationship between building sales and the year of build. While some building ID values exhibit statistical significance, the majority do not. Considering that building ID is a categorical variable, it is not among the most

valuable predictors for determining the sales amount. Upon comparison with other variables, gross square feet should be the most robust predictor due to its p-value which is way smaller than 0.05, indicating statistical significance and a strong relationship with building sales.

Part 2

To improve the precision of predictions and uncover patterns within sequential data, it is vital to employ time series forecasting techniques. Before constructing the prediction model, the sales data was converted into the appropriate date format, organized by year and quarter, the total sale prices were calculated for each quarter and arranged the data ascended based on the year and quarter. Subsequently, a quarterly time series was created using the sales data, commencing from the first quarter of 2009.

In order to determine the most appropriate model for forecasting sales, two different models were employed and compared. The first model applied the additive error, additive trend, and additive seasonality (AAA) model is appropriate for forecasting sales with consistent and predictable trends or seasonality. On the other hand, the additive error, non-additive trend, and additive seasonality (ANA) model is more suitable for capturing non-linear trends in the sales data, enabling a more comprehensive representation of complex patterns that offset the limitations of the linear trend. By comparing these two models, I can assess the ability of each model to accurately capture the historical patterns and trends exhibited in the sales data. Both prediction summaries are shown below in Figure1 and 2. I also computed the confidence band in Figure 3 which will be used in the optimization model in Part 3.

Comparing the two models, the ANA model shows a lower AICc and BIC and has a closer to 0 mean percentage error meaning the discrepancy of average percentage between

the forecasted values and the actual values in this model is smaller than the AAA model. Therefore, the ANA model may be more accurate in its prediction.

```
## ETS(A,Ad,A)
##
## Call:
##  ets(y = x_ts, model = "AAA")
##
##   Smoothing parameters:
##     alpha = 1e-04
##     beta  = 1e-04
##     gamma = 1e-04
##     phi   = 0.9575
##
##   Initial states:
##     l = 56974575.5451
##     b = 7095915.9221
##     s = 92190831 -30140123 -18744049 -43306659
##
##   sigma:  154508843
##
##      AIC     AICc      BIC
## 2176.582 2181.947 2196.094
##
## Training set error measures:
##                     ME      RMSE      MAE       MPE     MAPE      MASE
## Training set -3254775 140503097 88741828 -61.59844 85.05419 0.8204003
##                     ACF1
## Training set -0.1402491
```

Figure 1: AAA model summary

```
## ETS(A,N,A)
##
## Call:
##  ets(y = x_ts, model = "ANA")
##
##   Smoothing parameters:
##     alpha = 0.0906
##     gamma = 1e-04
##
##   Initial states:
##     l = 77972246.3285
##     s = 92190834 -30140123 -18744052 -43306659
##
##   sigma:  157485168
##
##      AIC     AICc      BIC
## 2176.073 2178.618 2189.732
##
## Training set error measures:
##                     ME      RMSE      MAE       MPE     MAPE      MASE
## Training set 19575902 148121091 96945293 -42.69592 88.78242 0.8962397
##                     ACF1
## Training set -0.1424035
```

Figure 2: ANA model summary

```
##         Point Forecast     Lo 80     Hi 80      Lo 95     Hi 95
## 2022 Q1      126874067 -74951296 328699431 -181791190 435539324
## 2022 Q2      151434168 -51217009 354085346 -158494062 461362399
## 2022 Q3      140043799 -63429841 343517439 -171142279 451229877
## 2022 Q4      262359096  58064498 466653693  -50082529 574800720
## 2023 Q1      126874067 -78236403 331984537 -186815326 440563460
## 2023 Q2      151434168 -54488941 357357278 -163498051 466366387
## 2023 Q3      140043799 -66688756 346776354 -176126360 456213958
## 2023 Q4      262359096  54818474 469899718  -55046895 579765086
```

Figure 3: ANA model future forecast with confidence band

Part 3

In addition to the joint and adjustment of data prior to the beginning of the analysis, I now execute the commercial building type data instead of filtering out the residential buildings to specifically analyze this part of the gain of opening a real estate office in my assigned neighborhood. By grouping the sales and years, I also calculated the average sale price per square foot with $515 per sqft in 2021 for further prediction.

To maximize the NPV of profit and develop an optimization model, additional assumptions were made. Firstly, the calculation is based on the assumption that the market penetration range, influenced by the commission, falls between a minimum of 4% and a maximum of 6%. The presence of employees was assumed to contribute to an additional increase in market penetration beyond the commission's influence. Secondly, the carry-over of the remaining budget from one quarter to the next was considered, facilitating comparison and potentially enhancing the NPV.

The model has applied GRG Nonlinear method since the model involves non-linear relationship and constraints. Assuming no budget carry-over in subsequent quarters, the estimated NPV from 2022Q1 to 2023Q4 is approximately $363.7 million, with 2 employees per quarter, a commission rate of 4.7%, and a market penetration of 7%. Conversely, with budget carry-over, the total NPV could reach around $376.3 million. However, changes in the number of employees are observed, with four employees required in 2022Q4, 2023Q2, and

2023Q4 due to higher sales and compounded budgets, while the remaining quarters only require one employee.

Considering the case of opening a real estate business in the HARLEM-CENTRAL neighborhood, hiring four employees bi-quarterly to achieve maximum profits may not be realistic. Adopting a more conservative approach, such as leaving the remaining budget unused, could mitigate risks associated with a newly established business. Therefore, it is advisable for the company to consider a commission rate of 4.7%, a market share of 7%, and employing 2 employees with no budget carry-over to optimize profits based on the predictions.