American University of Armenia

College of Science and Engineering

Capstone project

Queue prediction for multi-branch companies in the service industry

Author: Narine Isakhanyan
Supervisor: Aram Butavyan

Spring 2023

Abstract

The service sector is proliferating in Armenia, and many companies with various branches are dealing with the problem of queue management. The work proposes a model that, with specific changes, can be projected on an arbitrary product-service company. In the paper, the Coffee House Company was taken as an instance. Coffee House Company is Armenia's most significant and fastest-growing takeaway coffee shop network. Coffee House has 29 branches in Yerevan and regions. The menu of the Coffee House consists of more than 230 types of drinks. By developing four models and predicting future sales, the business can decide the number of baristas in advance to reduce and manage queues. It is demonstrated that additional information sources can improve the predictions' accuracy. This statement will be discussed in the following sections.

Contents

Chapter 1

Introduction

The data used for this research project was provided by Coffee House Company. Initially, the raw data was made up of thirty-nine monthly datasets; each dataset contained information about purchases of beverages from three branches in the period from January 2022 to January 2023. One of the branches is near the AUA, the second is near YSU, and the third is near RAU. Thus, for instance, the branch near AUA had thirteen datasets from January 2022 to January 2023. For the first step of the data exploration, all datasets were combined to understand the location where the purchase was made. A new variable University was created, which contained three objects: AUA, RAU, and YSU, which consists of 281340 entries. The data was cleaned, and any missing values were removed, which resulted in the number of observations being reduced to 269891. The raw data contained entries in multiple languages. Library from Google was used to translate most entries into English. Manual translations were required for some of the data.

It is crucial to predict the sales of any business, and coffee shops are no exception. Coffee shops are ubiquitous worldwide, and people periodically spend time standing in queues and purchasing beverages. Queues and long waiting times are a real struggle and common challenges for coffee shops, resulting in frustrated customers and lost revenues. Using the power of machine learning, it is possible to develop models that will help the company understand and predict where to allocate more employees to reduce the possibility of large queues. As Coffee House Company is the most famous coffee shop in Armenia, queues are also a problem for them. Thus, by predicting future sales, the management will know how to allocate baristas between branches to avoid queues. The project focuses on developing four machine-learning models for the Coffee House Company. The aim is to give the company powerful insights to manage their queues and enhance customer service. This research project thoroughly explores the dataset which The Coffee House Company provided. The dataset comes from three branches located near Yerevan State University (YSU), American University of Armenia (AUA), and Russian-Armenian University (RAU). It should be noted that each second order (observation) from the dataset has been removed before delivering the data. This means that before using the model, it can be improved by applying more data and more branches, in the same way, to draw better conclusions in terms of business.

The project is focused on a data-driven approach, as the most important and insightful information came from exploring the data and paying attention to understanding it. Therefore, the project is mainly based on two aspects.

1. Analysis of the sample data from the three branches of the Coffee House Company.
   a. Data Visualization.
   b. Feature extraction.
   c. Pattern detection.
2. The development of an initial time series model and four machine learning models, followed by best predictions:
   a. SARIMA model for seasonality detection.
   b. SARIMAX model using exogen variables.
   c. Linear Regression model.
   d. Poisson Regression model.
   e. Negative Binomial Regression model.

Chapter 2

Introduction to Data

The data used for this research project was provided by Coffee House Company. Initially, the raw data was made up of thirty-nine monthly datasets; each dataset contained information about purchases of beverages from three branches in the period from January 2022 to January 2023. One of the branches is near the AUA, the second is near YSU, and the third is near RAU. Thus, for instance, the branch near AUA had thirteen datasets from January 2022 to January 2023. For the first step of the data exploration, all datasets were combined to understand the location where the purchase was made. A new variable University was created, which contained three objects: AUA, RAU, and YSU, which consists of 281340 entries. The data was cleaned, and any missing values were removed, which resulted in the number of observations being reduced to 269891. The raw data contained entries in multiple languages. Library from Google was used to translate most entries into English. Manual translations were required for some of the data.
Table 2.1 shows the list of variables used for analysis.

Variable	Description of the Variable	Independent Variable
University	Near what institution the order was made.	✓
Date	The date the order was made.	✓
Open_Year	The year the purchase was done.	✓
Open_Month	The month the order was made.	✓
Open_Day_Month	The day the purchase was done.	✓
Open_Weekday	On which weekday the order was made.	✓
Service_Time	The seconds which took to prepare the order.	✓
Count	The number of orders made on that date.	×

Table 2.1: The variables University, Date, Open_Year, Open_Month, Open_Day_Month, Open_Weekday, Service_Time, and Count were used for analysis.

For feature engineering, we used four new datasets. The weather dataset describes the weather of the one year from 2022 to 2023 in Yerevan and also mentions the temperature. The subsequent datasets are about holidays and academic breaks at each university. A holiday dataset was created for each university, which comprised all national holidays and academic breaks. These datasets are the key factors to consider in sales prediction because special days of the year can impact sales. A hypothesis was made that the number of orders would likely decrease for these locations, on holidays, academic breaks, and weather conditions, given the abrupt schedule change.

2.1 Exploratory Data Analysis

The project's fundamental part is analyzing data to understand which models can be tested. The mean of total daily purchases of three branches is 656, which is very close to the average of AUA.

University	Count	Mean	std	min	25%	50%	75%	max	Total
AUA	122862	656	397	0.0	400	600	800	12000	80634331
RAU	90456	675	442	0.0	400	600	800	25200	61066846
YSU	56573	627	413	0.0	350	500	800	12500	35482586

Table 2.2: A thorough description of the branches near universities.

Table 2.2 shows measures for three branches: AUA, RAU, and YSU. There are several interesting findings:

- The average purchase varies slightly; however, there is no significant difference.
- The maximum purchase was made in the RAU; which is 25200 Armenian drams. The maximum purchase of AUA is 12000 drams, and the maximum at YSU is 12500 drams.
- There is a significant difference between the income and location of the Coffee House Company. The American University of Armenia generated 80 mln in income from January 2022 to January 2023. It has the highest income compared to RAU and YSU.
- Near AUA, almost two times more customers have been served than YSU.
- The median for AUA and RAU is the same; for the YSU, it is less.

To analyze the final price of sold goods, a distributional comparison per university has been made. Figure 2.1 shows the logarithmic final purchase price per university. It demonstrates that some outliers exist, but all values come from the same distribution.

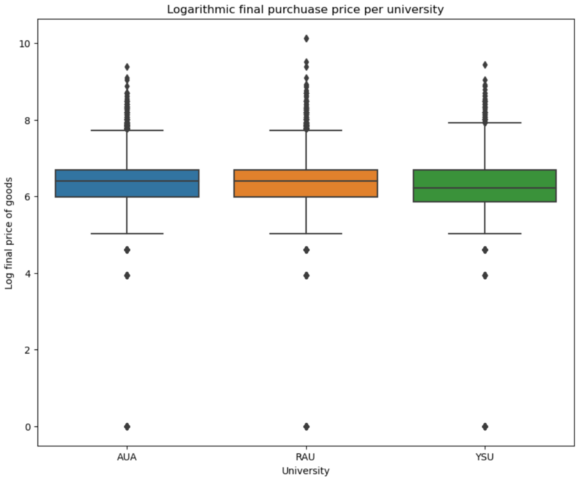


Figure 2.1: The logarithmic final purchase price for universities. X axis represents Universities and Y axis represents logarithmic price of goods.

Next thing that has been done is the exploration of customers per university and weekday.

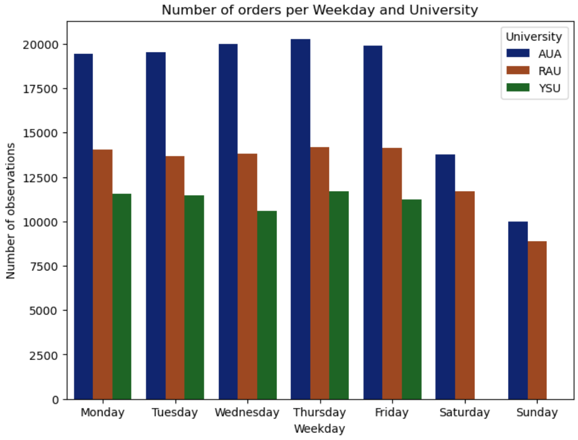


Figure 2.2: The number of orders per Weekday and University. AUA is in blue, RAU is in brown, YSU is in green. The number of observations is in the range from 0 to 20000.

Figure 2.2 shows that the number of orders is higher on weekdays because three branches of Coffee House are located near universities, and universities are active during weekdays. The decrease in orders is visible during weekends. Also, it should be noted that YSU is closed on Saturdays and Sundays. A similar comparison has been made using monthly splits of the data.

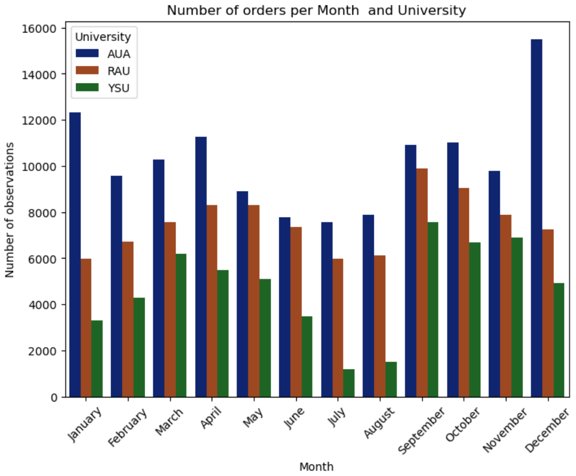


Figure 2.3: The number of orders per Month and University. AUA is in blue, RAU is in brown, YSU is in green. The number of observations is in the range from 0 to 16000.

Figure 2.3 shows that the number of orders decreased in the summertime. AUA has the highest number of orders compared to RAU and YSU. The reason is that AUA provides summer courses, and students visit the university. The academic community is shallow near YSU and RAU during summer as the next step distribution of orders per day has been analyzed. Figure 4 represents the distribution of orders amount per day across all universities.

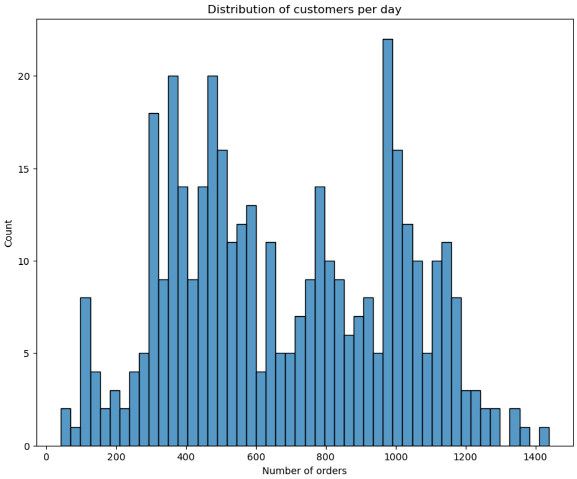


Figure 2.4: The distribution of customers per day. The number of orders is in the range from 0 to 1500.

From Figure 2.4, it can be concluded that the number of daily orders has a bimodal distribution. One in the range of 260-600 and another one from the range of 900-1200. To better understand the distribution, it is essential to understand the correlation between the number of orders and the total serving time of each university. To find hidden features in the data, the correlation between serving time and the number of orders has been calculated. Bellow in the heatmap is represented Pearson correlation results.

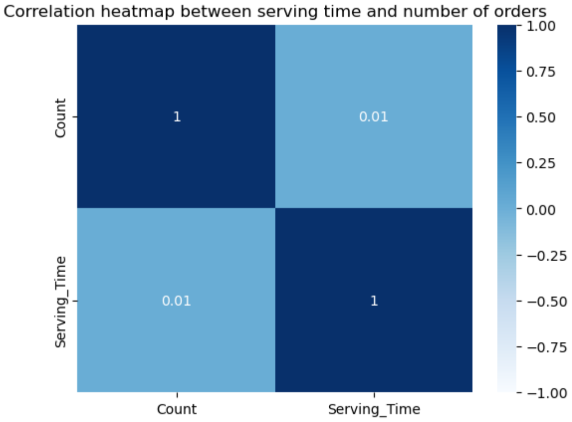


Figure 2.5: The correlation between serving time and the number of orders.

Figure 2.5 shows that the correlation coefficient is 0.01, which means that there is a weak relationship between serving time and the number of orders. Therefore, there is no need to add serving time to the model. Further analysis shows that there are distributional differences between the number of orders per university (see Figure 2.6).

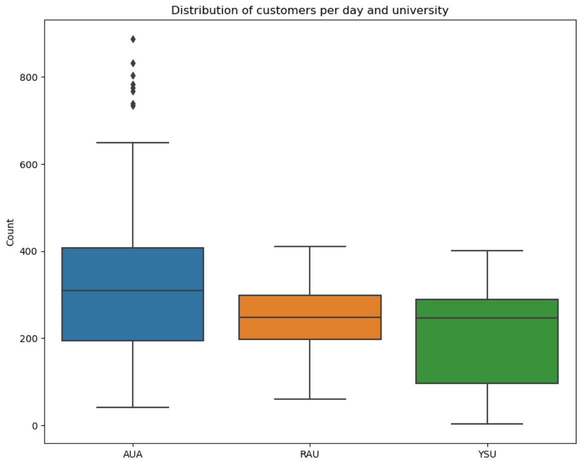


Figure 2.6: The distribution of customers per day and university. AUA is in blue, RAU is in orange and YSU is in green.

The number of daily orders was the most at AUA; moreover, there exist some outliers of American University of Armenia, which claim that there were days when the number of customers exceeded 800. A detailed description of daily orders at each branch is in Table 2.3. It maintains that the median for AUA is significantly higher than RAU and YSU. This means that the number of daily orders in AUA is considerably higher, as shown in Figure 2.6. In sharp contrast to this, the medians for RAU and YSU are very close to each other, 248 and 247, respectively. Also, the first quartile for AUA and RAU is almost the same, 195 and 197, respectively. For YSU, this measurement is significantly lower; it is 97. And finally, the table shows only at AUA 75% percentile exceeds 400.

University	Count	Mean	Std	Min	25%	50%	75%	Max
AUA	395	311	135	42	195	310	408	887
RAU	363	249	63	60	197	248	299	411
YSU	269	210	109	3	97	247	289	401

Table 2.3: Description of daily orders per university.

Chapter 3

Literature Review

According to Aiken L., generalized linear models, especially Poisson regression, bring accurate inference for non-clustered count data. One should integrate the multilevel models related to GLM models to understand the generalized linear mixed models [Aik+15]. For example, Seller K. points out that Poisson Regression is considered the most popular for modeling the count data. However, there are better ones to model accurate data. He introduced the COM-Poisson distribution, which is the Poisson distribution's generalization, allowing a large scale of overdispersion and under-dispersion. This new generalization gave flexibility for the fast growth of methodological research in many spheres [SBS12].

According to Sturman M., by comparing eight models (ordinary least squares, OLS with a transformed dependent variable, Tobit, Poisson, overdispersed Poisson, negative binomial, ordinal logistic, and ordinal probit regressions) for analyzing count data, they understood that every model could produce false positives. Nevertheless, the researchers expected that the OLS regression model would make many false positives, but it did not. Moreover, researchers revealed that every model has some false positives. Two models made false positives the most, which are the Tobit and Poisson regression models. Eventually, the Negative Binomial Model produced the least false positives [Stu99].

Shaon claimed that the traditional Negative Binomial Model could not handle highly over-dispersed data. He suggests using the Negative binomial Lindley model or the random parameters Negative Binomial model to make better modifications in the assumption of coefficients in data [Sha+18].

Maxwell argues in the research paper that Poisson distribution is the best for modeling count data. However, to work with a large scale of dispersion, it is better to use Negative Binomial Regression, Generalized Poisson Regression, Poisson Regression, and Conway-Maxwell-Poisson (COM-Poisson) Regression. This research paper compared the results with AIC and BIC, and the author declares that the Generalized Poisson regression performed the best [Max+18]. Also, Winkelmann claims that due to Poisson and Negative Binomial models, the Linear Regression Model will not support useful insights [Win15].

Additionally, Arunraj N. claims that it is beneficial to develop a SARIMA with external variables to forecast daily sales. Moreover, he states that the SARIMAX model enhances the conventional SARIMA model [AAF16].

Therefore, using Poisson Regression Model and Negative Binomial Model for the count data is better. However, it is also worth implementing Linear Regression Model for comparison with the abovementioned two models. Moreover, as the data is strictly related to time series analysis SARIMAX model will be applied in two steps. First, to detect seasonality, and second, to use exogen variables used for the Linear, Negative Binomial, and Poisson Regression to create a model to predict future order count. It should be noted that comparing all four models will be done using two statistical measurements: Akaike Informative Criterion and Mean Squared Error. This will let to choose the best model, which will be used for future prediction.

Chapter 4

Applied Models

Initially, the SARIMA model was developed on the dataset. Next, feature selection tests the correlation of each feature with the sales column and determines which features had the highest impact on sales. Finally, based on the results of the first two steps, four models were implemented: SARIMAX Model, Linear Regression Model, Poisson Regression Model, and Negative Binomial Regression Model. To select the best estimator, the dataset has been split into train and test parts, where the test dataset represents the last few days of observations. Such a split technique was conditioned by the type of data, which is time series. Additionally, the Akaike Information Criterion (AIC) has been used for model comparison in parallel with the mean squared error of the test data. The highest-performing model was selected as the predictive model, and prediction testing was performed. Besides, other statistical measures were computed to make comparisons between the models more accurate. These metrics will be discussed in the following sections. For the analysis, as a dependent variable, daily orders count has been used, and as independent variables, the following measurements have been used:

- University (two categorical variables, AUA, RAU, and YSU, where YSU is taken as a benchmark).

- Year (dummy variable with two unique values, where 2022 is taken as a benchmark)

- Day of week (dummy variables with six categories, where Monday is taken as a benchmark).

- Weather (dummy variable with seven categories, where Cloudy is taken as a benchmark).

- Holiday (dummy variable with two categories, where 1 is a holiday, 0 else).
- Also, one categorical feature was created to use the day of the month, showing if the day is in the first or second half of the month.

4.1 SARIMA Model

Well-known Autoregressive Integrated Moving (ARIMA) is considered the most actual method for time series and forecasting. SARIMA is the extension of ARIMA which can handle the trend of the data as well as the seasonality. The SARIMA model is the following:

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t \tag{4.1}$$

The SARIMA model contains an exceptional set of autoregressive and moving average components. SARIMA models allow for differencing data by seasonal frequency, yet also by non-seasonal differencing. Knowing which parameters are best can be made easier through automatic parameter search frameworks [Bro19]. SARIMAX model was also applied using same exogen variables that have been applied for Linear, Poisson and Negative Binomial Regression Models.

4.2 Linear Regression Model (OLS)

$$y = \beta_0 + \beta_1 X + \beta_2 X + \cdots + \beta_k X + \varepsilon \tag{4.2}$$

The most common technique for linear regression is the Ordinary Least Square method. By using ordinary least squares regression, it is possible to estimate coefficients of linear regression equations, which explains the relationship between dependent and independent variables [Kum23]. Model coefficients were computed using following formula [Unknd]:

$$\hat{\beta} = \frac{cov(x, y}{Var(x)} \tag{4.3}$$

To understand if the parameters are significant following hypothesis has been used [Abdnd]:

$$H_0 : \beta = 0 \tag{4.4}$$

$$H_A : \beta \neq 0 \tag{4.5}$$

All the tests have been computed for a 95 % confidence interval. In the analysis results description only, significant parameters will be reported.

4.3 Poisson Regression Model

Poisson distribution is a discrete probability distribution. When a given number of events happens in a fixed interval of time or space and has a constant mean rate, the Poisson distribution describes its probability. General formula for the model is following [Ber22]:

$$\lambda = e^{\beta_0 + \beta_1 X + \beta_2 X + ... + \beta_k X} \tag{4.6}$$

The probability of observing x events within a given interval is calculated with the following formula [Al,nd]:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4.7}$$

One of the most essential properties of Poisson distribution and Regression is equidispersion, which will be discussed in the analysis section.

4.4 Negative Binomial Regression Model

The generalization of the Poisson regression is known as Negative Binomial regression. The promotion of the Negative Binomial that it has a restrictive hypothesis that the variance and the mean made by the Poisson model are equal. The traditional negative binomial regression model is founded on the Poisson-gamma combination distribution. This formulation is widespread because it permits the modeling of Poisson heterogeneity using a Gamma distribution [P.21]. In Negative Binomial regression, the mean of y is decided by the exposure time t and a set of k regressor variables (the x's). The expression relating to these quantities is [NCSnd]:

$$\mu_i = e^{lnt_i + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k x_{ki}} \tag{4.8}$$

The observation i of the Negative Binomial Regression is written like this [NCSnd]:

$$Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1}\Gamma(y_i + 1))} \times (\frac{1}{1 + \alpha\mu_i})^{\alpha^{-1}} \times (\frac{\alpha\mu_i}{1 + \alpha\mu_i})^{y_i} \tag{4.9}$$

Chapter 5

Results

5.1 SARIMA Model Results

As mentioned in the Model Development section SARIMA Model has been applied to determine whether any type of seasonality is present in the data. Figure 5.1 is the time series plot representing the number of daily orders.

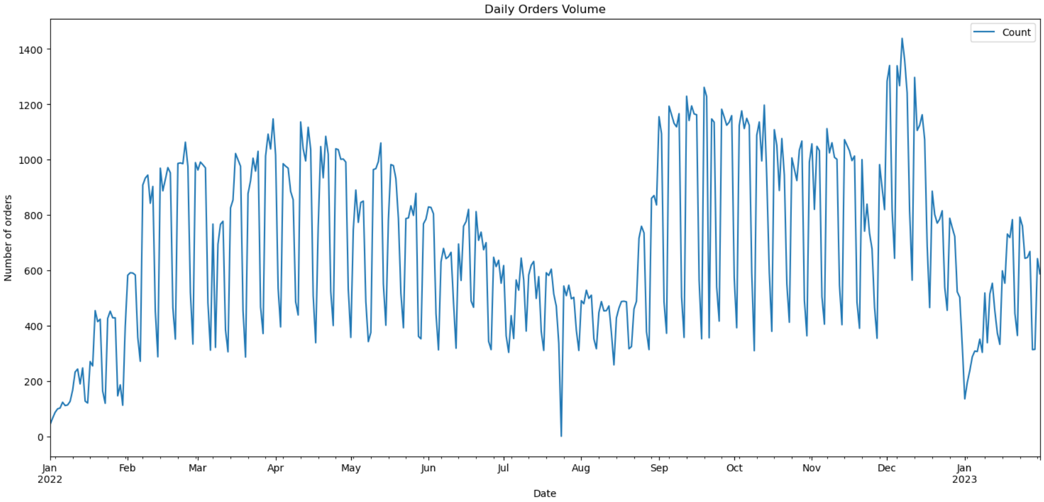


Figure 5.1: The daily orders volume. The number of orders is in the range from 0 to 1600. The date is in the range from January 2022 to January 2023.

Before creating the model, time series decomposition was derived. In Figure 5.2 there are presented the results of the decomposition.

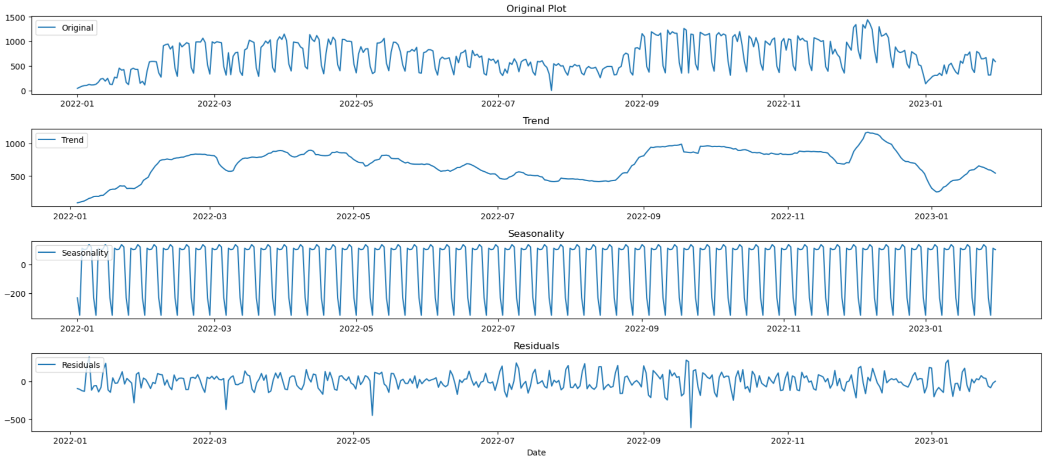


Figure 5.2: The trend, seasonality, and residual decomposition.

From the figure above (Figure 5.2) we can conclude that there exists seasonality, but there is no vividly expressed trend. It should be noted that before model development, the Dickey-Fuller test was used to check the stationarity of the data, where the null hypothesis is the absence of stationarity. At a 95 % confidence interval, the null hypothesis was rejected as the p-value of the test was equal to 0.02. To understand which lags to take, we build the autocorrelation and the partial autocorrelations figures, presented below.

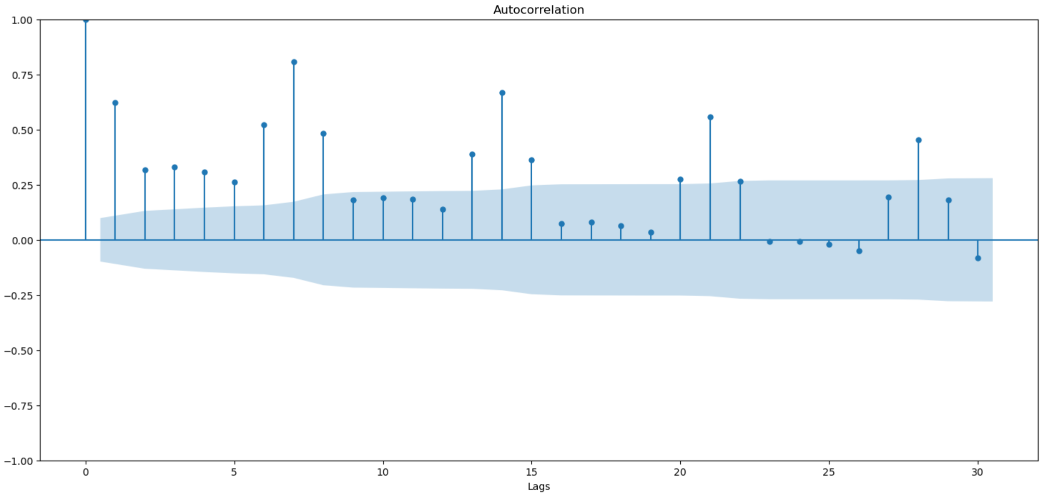


Figure 5.3: Autocorrelation and lags.

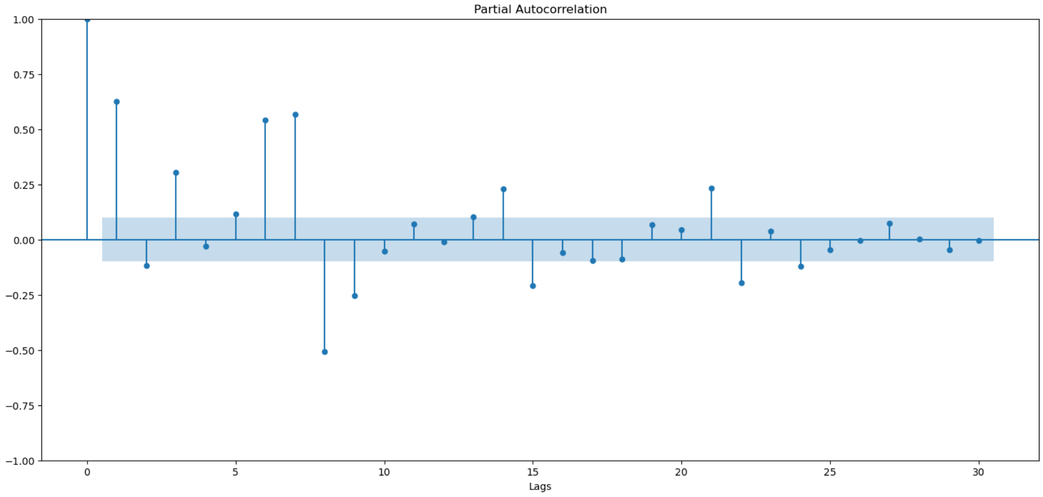


Figure 5.4: Partial Autocorrelation and lags

Autocorrelation (Figure 5.3, 5.4) shows that there is weekly seasonality, thus the seasonality is 7. Due to the Grid Search, the SARIMA parameters have been derived, as well as the best model. After the construction of the model, the best model was chosen ARIMA (1, 1, 2) x (1, 1, 2, 7), and the Akaike Information Criterion is 4666.59.
Figure 5.5 is the representation of residual diagnostics; the conclusion is that it is almost normally distributed.

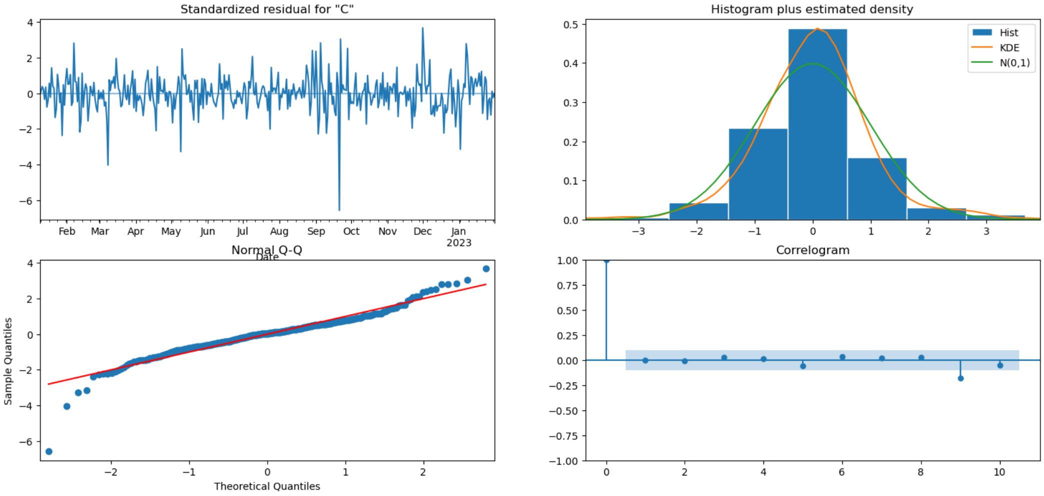


Figure 5.5: Top Left: Standardized residuals for “C”. Top Right: Histogram plus estimated density. . Bottom Left: Normal distribution. Bottom Right: Correlogram.

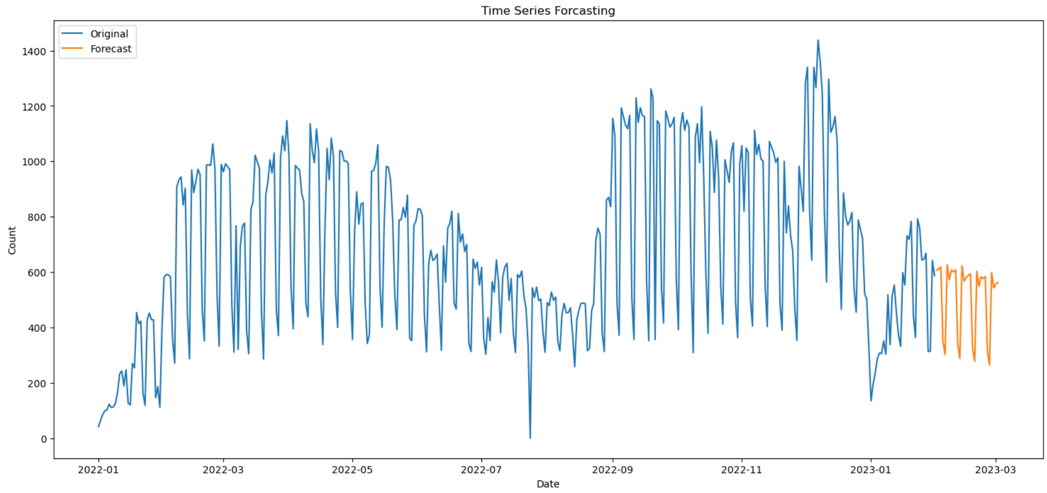
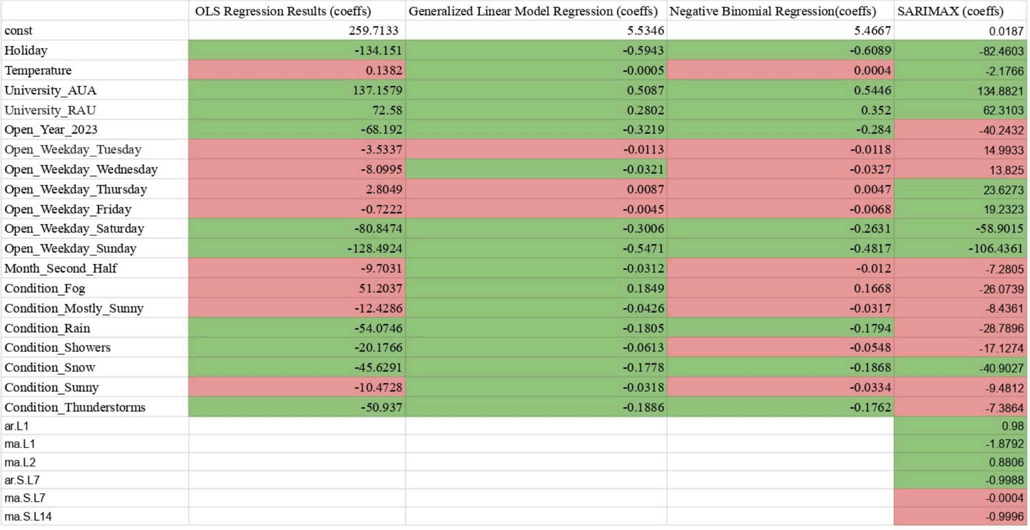


Figure 5.6: Time Series Forecasting. The original distribution is in blue, the Forecasted is in Orange.

The following figure (Figure 5.6) represents the forecast for the number of customers until March 2023 based on the SARIMA Model.

5.2 Summary of the Results of four models

In table 5.1 it is presented combined results of the developed models. Details of the models' output will be discussed in the next sections.



	OLS Regression Results (coeffs)	Generalized Linear Model Regression (coeffs)	Negative Binomial Regression(coeffs)	SARIMAX (coeffs)
const	259.7133	5.5346	5.4667	0.0187
Holiday	-134.151	-0.5943	-0.6089	-82.4603
Temperature	0.1382	-0.0005	0.0004	-2.1766
University_AUA	137.1579	0.5087	0.5446	134.8821
University_RAU	72.58	0.2802	0.352	62.3103
Open_Year_2023	-68.192	-0.3219	-0.284	-40.2432
Open_Weekday_Tuesday	-3.5337	-0.0113	-0.0118	14.9933
Open_Weekday_Wednesday	-8.0995	-0.0321	-0.0327	13.825
Open_Weekday_Thursday	2.8049	0.0087	0.0047	23.6273
Open_Weekday_Friday	-0.7222	-0.0045	-0.0068	19.2323
Open_Weekday_Saturday	-80.8474	-0.3006	-0.2631	-58.9015
Open_Weekday_Sunday	-128.4924	-0.5471	-0.4817	-106.4361
Month_Second_Half	-9.7031	-0.0312	-0.012	-7.2805
Condition_Fog	51.2037	0.1849	0.1668	-26.0739
Condition_Mostly_Sunny	-12.4286	-0.0426	-0.0317	-8.4361
Condition_Rain	-54.0746	-0.1805	-0.1794	-28.7896
Condition_Showers	-20.1766	-0.0613	-0.0548	-17.1274
Condition_Snow	-45.6291	-0.1778	-0.1868	-40.9027
Condition_Sunny	-10.4728	-0.0318	-0.0334	-9.4812
Condition_Thunderstorms	-50.937	-0.1886	-0.1762	-7.3864
ar.L1				0.98
ma.L1				-1.8792
ma.L2				0.8806
ar.S.L7				-0.9988
ma.S.L7				-0.0004
ma.S.L14				-0.9996

Table 5.1: Results of Linear Regression Model, Generalized Linear Model Regression, Negative Binomial Regression and SARIMAX Model with coefficients. Note: the green cells represent significant coefficients, the coral cells represent non-significant coefficients.

5.2.1 SARIMAX Model

The first model that has been used is SARIMAX Model. From the results (Table 5.1) of the model, it can be concluded that:

- Holding all other variables constant during Holidays, the number of orders on average decreases by 82.
- Holding all other variables constant, if the temperature decreases by 1 unit of Celsius, the number of orders on average decreases by 2.
- Compared with AUA and RAU, the number of orders in YSU on average is less by 134 and 62.
- There is no significant difference between Monday, Tuesday, and Wednesday. However on Thursday and Friday, the number of orders increases, and sharply decreases on weekends.

The AIC is 11228.3 and the Mean Squared Error is 3794.48. In Figure 5.7 is represented prediction and the actual amount of orders line graph.

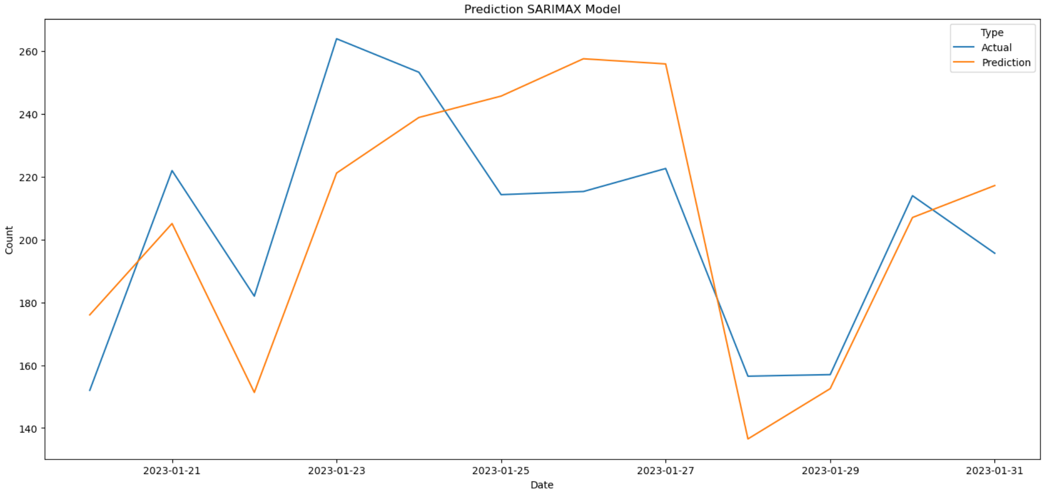


Figure 5.7: The prediction of SARIMAX model. The original distribution is in blue, and the prediction is represented in orange. The prediction range is from 100 up to 300.

5.2.2 Linear Regression Model

The second model that has been developed is Linear Regression with the OLS method. From the results (Table 5.1) of the model following conclusions have been made:

- Holding all other variables constant in RAU and AUA in average daily served customers are more than in YSU by 72.6 and 137.2 respectively.

- Holding all other variables constant, sales in 2023 decreased on average by 68.2.

- In terms of sales, there is no significant difference between weekdays; however, sales sharply decrease at weekends.

- R2 is equal to 54.5 %. This means that from all orders 54.5% of variance is explained.

- The Jarque Bera hypothesis test shows that the normality test for residuals is rejected.

- When there is Rain, Showers, Snow and Thunderstorms the sales sharply decrease by 54, 20.1, 45.6, 50.9 times respectively compared to Cloudy weather.
- The temperature is non-significant, the assumption is that during cold days people order warm beverages.

Also, we got that the AIC is 11539.7 and the Mean Squared Error is 3559.03. Figure 5.8 is the prediction of the Linear Model:

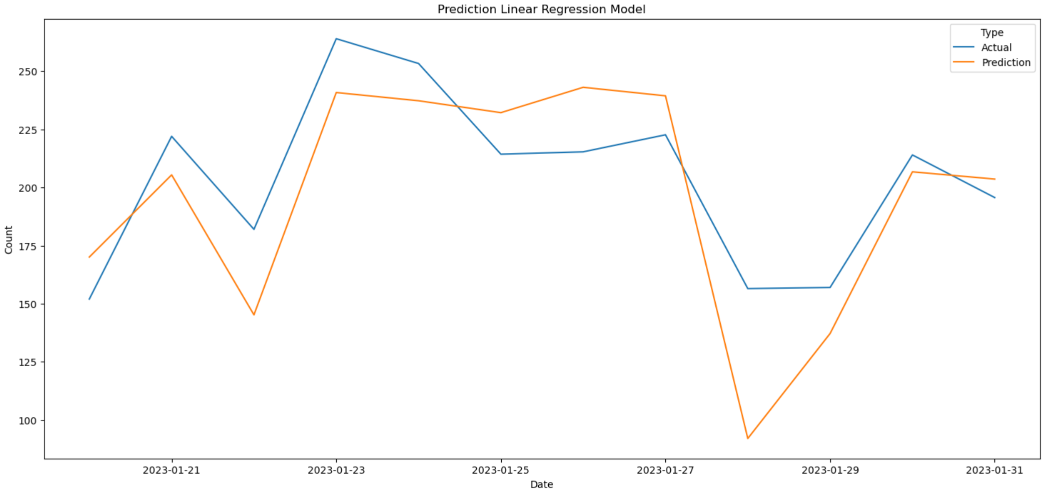


Figure 5.8: The prediction of Linear Regression Model. The original distribution is in blue, the Forecasted is in Orange. The range of Date is from the beginning of 2022 to the end of January 2023. The dependent variable Count is in the range from 100 to 300.

5.2.3 Poisson Regression Model

The third model that has been used is Poisson Regression Model, which is mostly used for count data estimation.

From the results (Table 5.1), it is evident that:

- Holding all other variables constant in RAU and AUA, average daily served customers are more than in YSU by 1.32 and 1.66 times, respectively. exponential
- Holding all other variables, constant sales in 2023 decreased on average 1.38 times.
- The pattern related to weekdays is the same as for linear regression.

- During the second half of the month, the number of sales decreased by 1.03, and this coefficient is significant at 95 %. It should be noted that the days during the second half are fewer than in the first half due to February.

- Weekdays are significant, only Tuesday, Thursday and Friday are non-significant.

- Each kind of weather is significant.

- The temperature is slightly significant.

Also, we got that the AIC is 29613.3 and the Mean Squared Error is 3938.6. The prediction of the model is the following (see Figure 5.9):

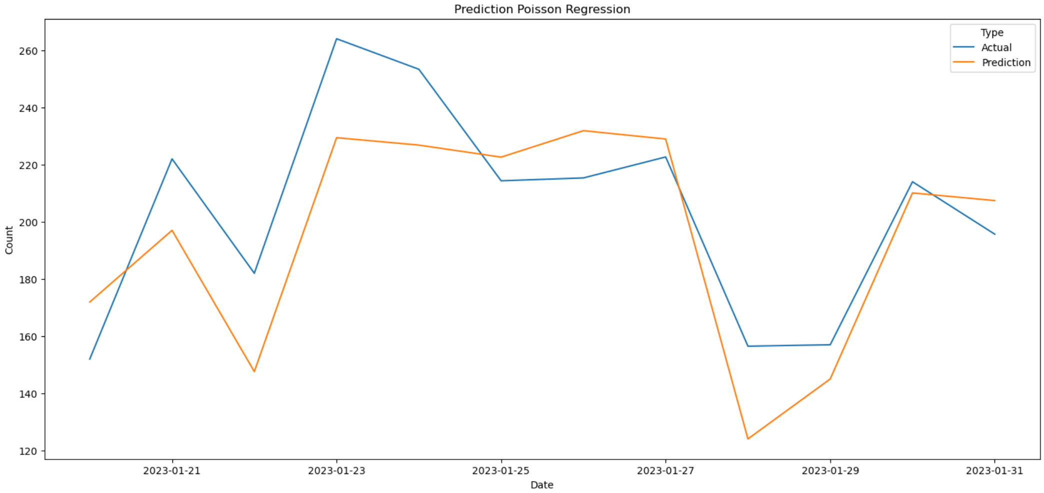


Figure 5.9: The prediction of Poisson Regression Model. The original distribution is in blue, the Forecasted is in Orange. The range of Date is from the beginning of 2022 to the end of January 2023. The dependent variable Count is in the range from 120 to 300.

5.2.4 Negative Binomial Regression Model

The final model that has been tested was the Negative Binomial Model.

From the results (Table 5.1), it is evident that:

- Holding all other variables constant in RAU and AUA, average daily served customers are more than in YSU by 1.42 and 1.72 times, respectively.

- Holding all other variables constant, sales in 2023 decreased on average 1.33 times (very close to Poisson regression model results)
- Pattern related to weekdays is very close as for linear and Poisson regression.
- The weather is slightly significant.
- Sales sharply decrease during Rain, Snow and Thunderstorms 1.2, 1.2 and 1.19 times respectively compared with cloudy.
- The temperature is non-significant, the assumption again is that people start to buy warm beverages during the cold times.

Also, we got that the AIC is 11638.9 and the MSE is 3650.2. Figure 5.10 is the representation of the prediction:

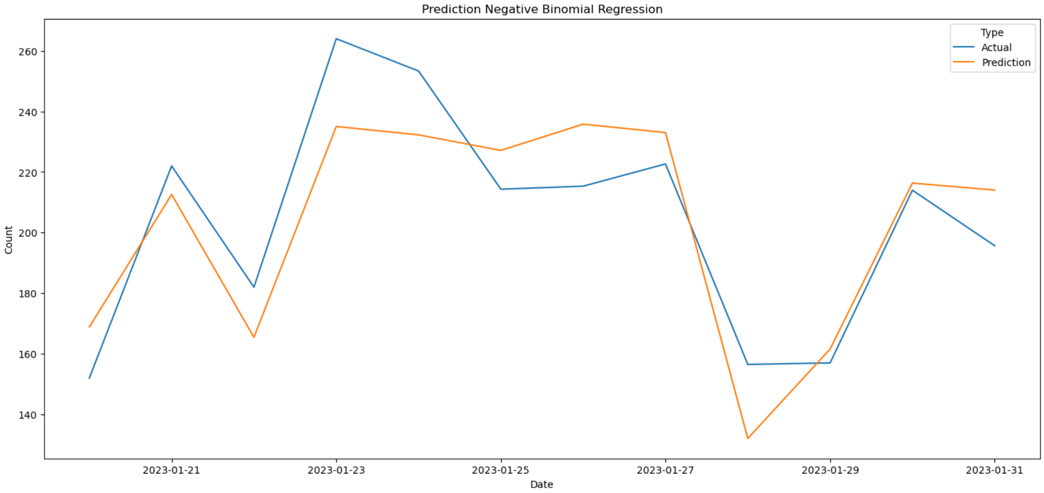


Figure 5.10: The prediction of the Negative Binomial Regression Model. The original distribution is in blue, the Forecasted is in Orange. The range of Date is from the beginning of 2022 to the end of January 2023. The dependent variable Count is in the range from 140 to 300.

5.3 Comparison of results and prediction

The models gave us actual results (see Table 5.2). For example, from the table, one can claim that the Akaike Information Criterion of Linear and Negative Binomial Regression Models are very close. However, the AIC of the Poisson Regression Model is equal to 29613.3, which concludes that the Poisson Regression Model is the worst model of these three models. The smallest Mean Squared Error has a Linear Regression Model. Nevertheless, the Linear Regression Model has the smallest

	SARIMAX Model	Linear Reg. Model	Poisson Reg. Model	Neg. Bin. Reg. Model
AIC	11228.3	11539.7	29613.3	11638.9
MSE	3794.5	3559.03	3938.6	3650.2

Table 5.2: Akaike Information Criterion and Mean Squared Error results of SARIMAX Model, Linear Regression Model, Generalized Linear Model Regression and Negative Binomial Regression.

AIC and MSE (see Table 5.2); however, as mentioned above, the residuals are not normally distributed. Moreover, we have count data; thus, we will take the best model for our dataset Negative Binomial Regression Model. All predictions, although they show some difference in the predicted values, present the correct pattern of trends for Coffee Houses, which means that using any above-suggested model will demonstrate where to expect more customers and the number of orders, so decisions will be made correctly in terms of where more employees are required. The best model, the Negative Binomial Model, prediction for the future number of orders has been calculated for the upcoming few weeks to help effectively organize the management of queues.

Figure 5.11 are represented the results of the prediction.

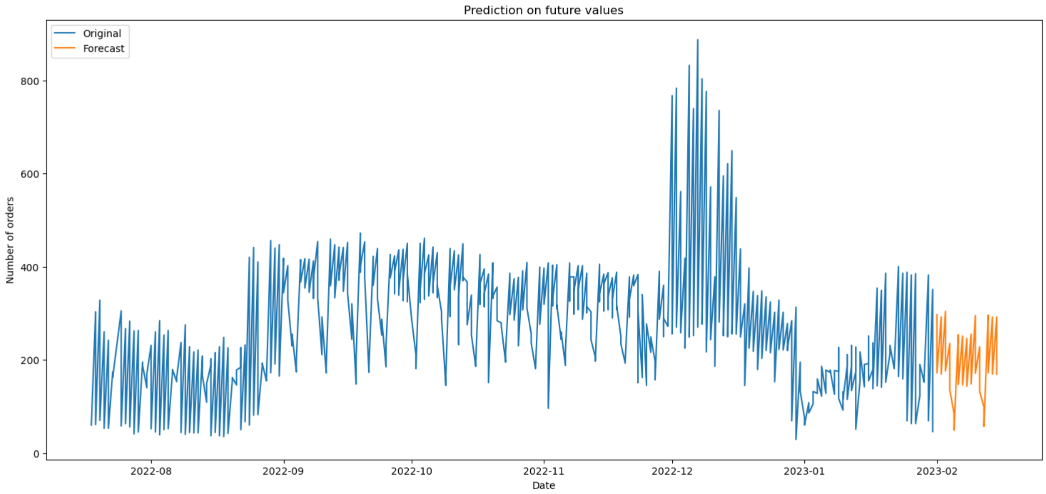


Figure 5.11: Negative Binomial Model Prediction. The prediction of the future values the original distribution is in blue, the Forecasted is in Orange. The range of Date is from the beginning of 2022 to the middle of February 2023. The number of orders is in the range from 0 to 900.

Chapter 6

Conclusion

The project focused on analyzing queue prediction for multi-branch companies in the service industry. As an example, Coffee House private data has been used for model development, which has never been used before for such type of analysis. Four models have been tested: SARIMAX, Linear, Poisson, and Negative Binomial Regression. A comparison of the models has been made using the mean squared error of test data and the Akaike Informative Criterion. Although the mean squared error and AIC of Linear regression have been slightly better than Negative Binomial Regression model estimations, residuals of the Linear Regression were not normally distributed, also considering the literature review, the Negative Binomial Regression model has been selected as the best estimator. It is essential to note that the developed model can be projected to any company that will provide data for the analysis as the main features are general (same) for all similar companies.

Bibliography

[Stu99] Michael C. Sturman. “Multiple Approaches to Analyzing Count Data in Studies of Individual Differences: The Propensity for Type I Errors, Illustrated with the Case of Absenteeism Prediction”. In: *Educational and Psychological Measurement* 59.3 (1999), pp. 414–430.

[SBS12] Kimberly F Sellers, Sharad Borle, and Galit Shmueli. “The COM-Poisson model for count data: a survey of methods and applications”. In: *Applied Stochastic Models in Business and Industry* 28.2 (2012), pp. 104–116.

[Aik+15] Leona S Aiken et al. “Analyzing count variables in individuals and groups: Single level and multilevel models”. In: *Group Processes & Intergroup Relations* 18.3 (2015), pp. 290–314.

[Win15] Rainer Winkelmann. “Quantitative policy evaluation can benefit from a rich set of econometric methods for analyzing count data”. In: *wol.iza.org* (2015).

[AAF16] Nisha Arunraj, Daniel Ahrens, and Marcus Fernandes. “Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry”. In: *International Journal of Operations Research and Information Systems* 7.2 (2016), pp. 1–21.

[Max+18] O. Maxwell et al. “Modelling Count Data; A Generalized Linear Model Framework”. In: *American Journal of Mathematical and Statistical Research* 6.3 (2018), pp. 55–63.

[Sha+18] Mohammad R.R. Shaon et al. “Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly overdispersed crash count data”. In: *Analytic Methods in Accident Research* 18 (2018), pp. 33–44.

[Bro19] Jason Brownlee. *A Gentle Introduction to SARIMA for Time Series Forecasting in Python*. 2019. URL: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/ (visited on 05/05/2022).

[P.21] Team P. *Poisson Regression and Negative Binomial Regression - Algoritma Data Science School*. https://algoritmaonline.com/poisson-regression-and-neg-ative-binomial-regression/. 2021.

[Ber22] S. M. van den Berg. *Analyzing data using linear models. Chapter 16 Generalized linear models for count data: Poisson regression*. https://bookdown.org/pingapang9/linear_models_bookdown/poisson.html. 2022.

[Kum23] A. Kumar. *Ordinary Least Squares Method: Concepts & Examples - Data Analytics*. https://vitalflux.com/ordinary-least-squares-method-concepts-examples/. 2023.

[Abdnd] Hervé Abdi. *Chapter 9 simple linear regression - Carnegie Mellon University*. https://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf. n.d.

[Al,nd] Al, B. I. & O. E. *Poisson Distribution | Introduction to Statistics*. Retrieved April 30, 2023, from https://courses.lumenlearning.com/introstats1/chapter/poisson-distribution/. n.d.

[NCSnd] NCSS. “Negative Binomial Regression”. In: (n.d.).

[Unknd] Unknown. *The First Method for Finding*
$beta_0$ *and*
$beta_1$. Online. n.d.