ALCOHOL PROJECT

February 1,2023

Introduction

We are a data consulting company which provide analyses for food and beverage market. We are currently operating two projects regardings alcohol data.

WINE QUALITY PROJECT

1 Problem statement

An Italian Wine Trading Company would like to evaluate wine quality from their suppliers before the acquisition. They send us the list of wine's composed chemicals. On the first project we tried to define Legit and Fraud wine based on these information.

2 Dataset

The used dataset includes 6497 observations with 13 variables with no missing or duplicates. The wine fraud dataset, described in Table 1, contains information about the chemical composition of various wines. At first, we conducted an Exploratory Data Analysis on the whole dataset to have an idea about its structure and contents. Afterward, we also calculated the correlation between variables and displayed some plots (Figure 1 and 2) of the most relevant variables.

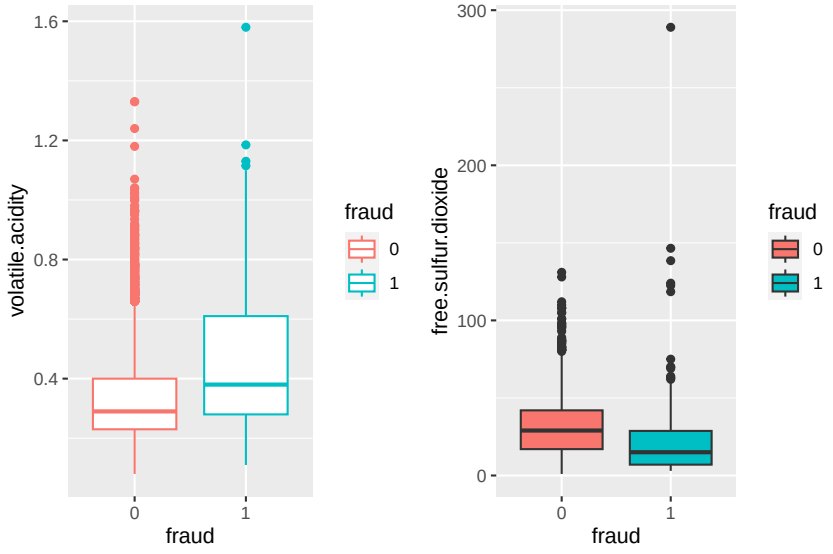
Considering the purpose of the analysis, we converted "quality" and "type" variables from characters into numbers. Then, we randomly divided the available sample into training set *(70% of the dataset)*, used to estimate the mode, and testing set *(the remaining 30%)*, used to evaluate the predictive accuracy of the model. The target variable is **Fraud**, assigning one to fake wine and zero otherwise. After several considerations on the importance of the variables included in the dataset for the predictability of Fraud, we decided to focus on *volatile acidity, residual sugar, free sulfur dioxide, density* and *red* (assigning one to red wine and zero to white wine).

Table 1: Variable descriptions

Variable	Description	Type
fixed acidity	fixed acids involved with wine	Numeric
volatile	the amount of acetic acid in wine	Numeric
citric acid	citric acid	Numeric
residual sugar	the amount of sugar remaining	Numeric
chlorides	the amount of salt in the wine	Numeric
free sulfur dioxide	the free form of SO2	Numeric
total sulfur dioxide	amount of free and bound forms of S02	Numeric
density	the density of water	Numeric
pH	scale from 0 (very acidic) to 14 (very basic)	Numeric
sulphates	sulfur dioxide gas (S02) levels	Numeric
alcohol	the percent alcohol content of the wine	Numeric
red	Red wine or white wine	Character
quality	Wine quality	Character

Since we also considered "red" variable can be used to estimate the quality of wine, the Person's Chi-square test is required to check the dependencies between the categorical and the target variable. A Chi-square test does not reject the null hypothesis of independence *(p-value = 0.7679)*.

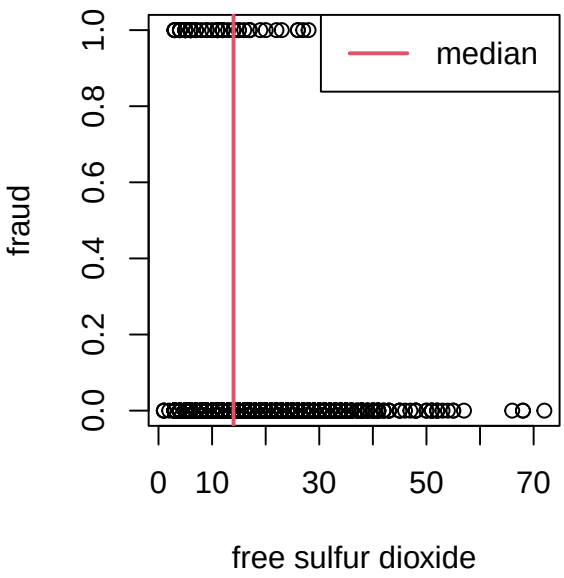
Figure 1: Relationship between Fraud and volatile acidity/free.sulfur.dioxide.



3 Method

Based on the goal of the company, we decided to build an algorithm predicting the probability of Fraud of a precise wine and then classify it as Fraud or not. As the dependent variable Fraud can only take two possible outcomes. We assumed that Y_i = Fraud is a random variable that follows a Bernoulli distribution, hence $Y_i = 1$ if the ith wine is Fraud

Figure 2: Median plot



($p_i > 0.5$) and $Y_i = 0$ otherwise, the probability of Fraud for the i-th wine is given by the conditional expected value $E(Y_i|X_i) = p_i$, where X_i =(*volatile acidity, residual sugar, free sulfur dioxide, density , red*) and $p_i \in (0, 1)$.

4 Analysis

Recall the testing set to access the accuracy of the classification using the confusion matrix and the ROC curve.

We obtained a large accuracy rate about **96.3%** for the Short-model, which is almost the same rate as we predict all the cases labelled as 0 (**96,1%**). In other words this high rate is not due to the quality of the model but rather due to the imbalanced classes. Indeed, if we look at the specificity rate, it is about **3.9%** indicating that the model poorly predict the fraudulent wine which is the most important class label that we want to predict correctly. Again, the Area under the curve is about **52%**, confirming that the poorly fit of the model (Figure 3).

The main reason for this misleading result is due to the imbalance in the original dataset, where the percentage of fraudulent wine is only **3.8%** of the whole dataset. To overcome this problem, we carried out some common subsampling methods such as adjusting sampling weight, up sampling, down sampling and ROSE on the training set.

The accuracy score, specificity rate of each method are showed on the Table 2, while the Area under the curve is drawn on Figure 4.

Figure 3: ROC curve of original dataset

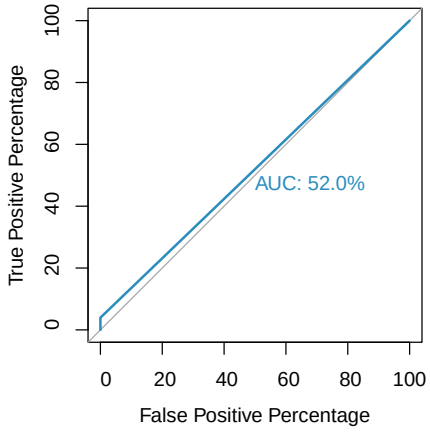
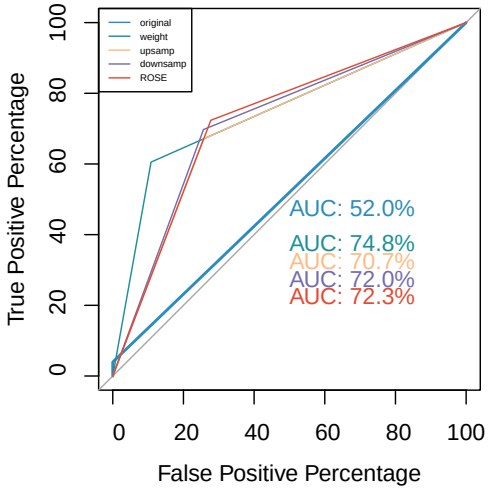


Table 2: Accuracy rate, Specificity rate, and Area under the curve for each method

	Accuracy score	**Specificity**	**AUC**
Original dataset	96.3%	3.9%	52.0%
Adjusted weight	88.1%	60.5%	74.8%
Up sampling	74.1%	67.1%	70.7%
Down sampling	74.2%	69.7%	72.0%
ROSE	72.3%	72.4%	72.3%

Figure 4: ROC curve



Referring to the obtained result, we would apply ROSE method to solve imbalance problem. Although ROC curve shows its AUC = **72.3%**, which is smaller than adjusted weight sample (**74.8%**), it has the highest Specificity rate (**74.2%**), predict the fraudulent wine which is the most important class label that we want to predict correctly.

Even we applied many methods to overcome the imbalance problem of this dataset,

the AUC did not improve that much. We can conclude that, the chemical ingredients are not the good predictors to verify the quality of wine. We would suggest the company collect more data such as bottle, label Investment firms etc., since the Legit and Fraud wine may also have very similar proportion of composed chemicals.

HY-VEE SALE PROJECT

5 Problem Statement

Hy-Vee, Inc. is one of the most impactful chain of supermarkets in the Midwestern and Southern United States, with more than 280 locations in Iowa, Illinois, Kansas and so on. Hy-Vee stores are full-service supermarkets with bakeries, delicatessens, floral departments, dine-in and carryout food service, wine and spirits, pharmacies, health clinics,... This project is carried out based on the requirement of Iowa Department of Commerce, Alcoholic Beverages Division, who wish to know the forecast of future alcohol sale from all Hy-vee stores in Iowa state.

6 Dataset

The dataset is dowloaded from State of Iowa's public data platform, which contains the purchase information of Iowa Class "E" liquor licensees by product sold from grocery stores, liquor stores, convenience stores, etc. across Iowa State. We have processed the original dataset and extracted only daily sale (in dollars) between the period of 2019 and 2021 from all the stores, supermarkets under Hy-vee brand. Only sales during five of days of a week were recored and sale record for weekends were originally left out. As a result, we obtained a dataset with 754 observations with neither NAs nor duplicated rows. The dataset was later split into a training set *(70% of total observation)* and is used to fit different modes while the testing set *(30% remaining of the observations)* is used to evaluate the performance of those models. Consequently, we generate time series of daily sale and decompose such time series into different components which includes trend, seasonality and residuals. This way we obtained the plots in figure 5 and 6.

Figure 5: Sale over the period of time from 2019 to 2021

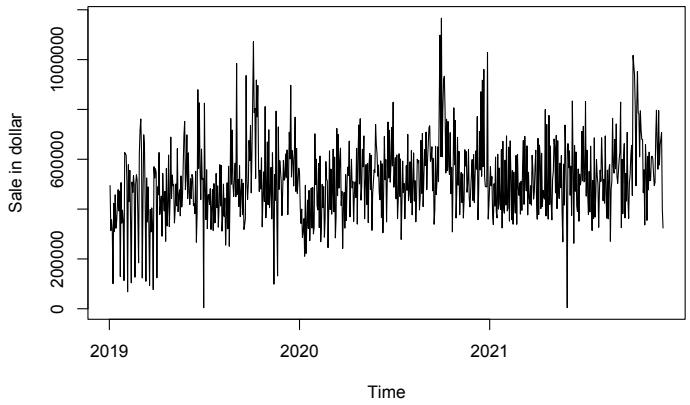


Figure 6: Decomposition of Sale time series

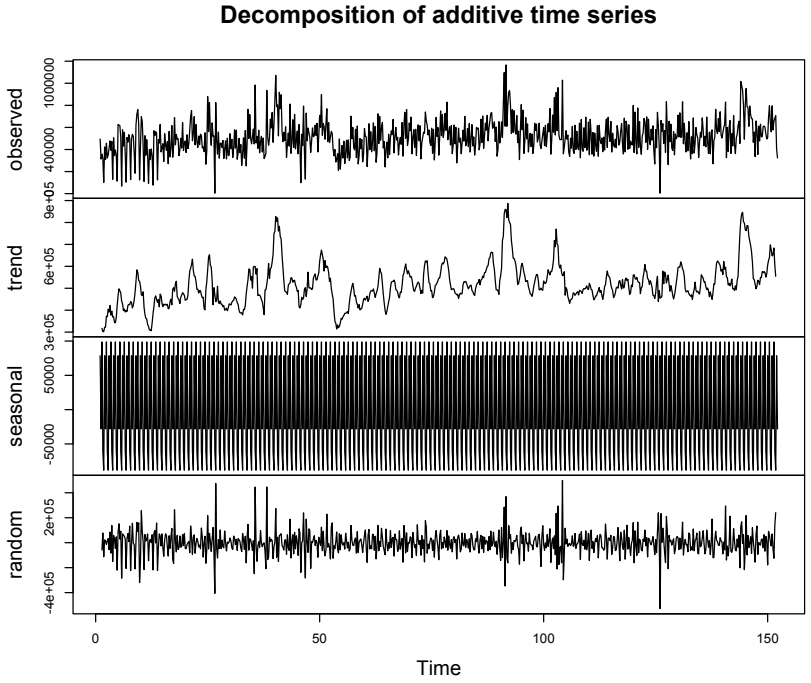
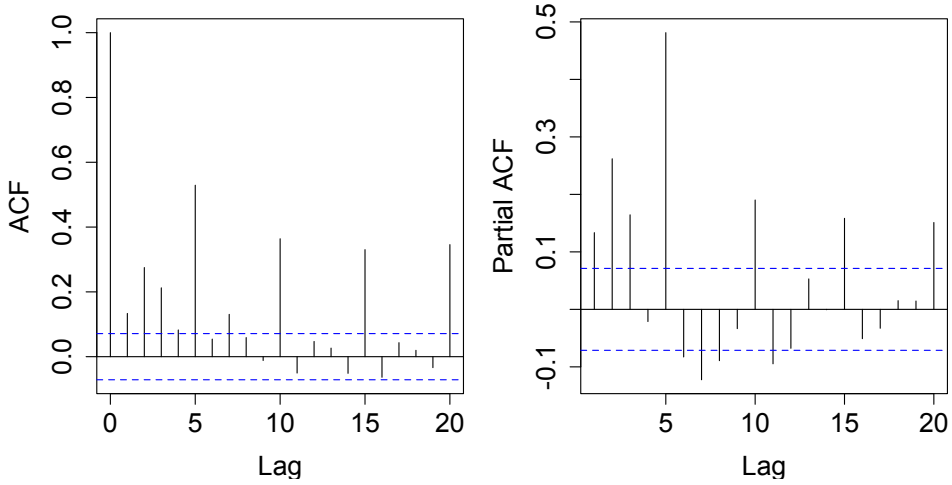


Figure 7: ACF and PACF of sale time series



From figure 6 we can suggest that a clear trend is not strongly present in our time series which implies the data is stationary. To prove this initial intuition of stationarity (no unit root), we performed Augmented Dickey-Fuller test and attain the result with p-value = 0.01, which reject the null hypothesis of non-stationarity. Followingly, we created ACF and PACF plots of sale time series to detect the autocorrelation between the time series and its own lagged (Figure 7).

The ACF plot does not have exponentially decaying behaviour due to the fact that there is no clear trend in the dataset. There are nine significant lags out of twenties, which proves that there exists the correlation between the current value and the value in previous periods; thus, this dataset is not a random walk. Similarly with the PACF plot, there are ten significant lags out of twenties.

7 Method

Now we proceed to build a model that can predict well the future sale of alcoholic beverage of Hy-vee stores. For this purpose, we have fitted the training set of the sample with various models to see which one returns best prediction and the metric we use to evaluate the performance of a model are RMSE and AIC. Firstly, we tried to fit t linear and quantile regression on sales: both performed decently well. Quantile regression was able to catch an impressive amount of true values inside the quantile range (according the prediction plot in R script). This result is not unforeseen since the linear regression models (quantiles included) are expected work quite well on stationary time series which is the case of our dataset. However, forecasting time series is always a challenge task and a linear line is not always considered as the most optimal solution when forecasting the future values, we took a further step considering an ARMA model which can predict future values based on past values. As our data is already stationary so the differencing step is not necessary

indicating that d parameter in ARIMA equals to 0; thus; we will use ARMA model to forecasting the future values of sales.

$$ARMA(p,q) : Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-1} + \sum_{i=1}^{q} \theta_i \epsilon_{t-1} + \epsilon_t$$

8 Analysis

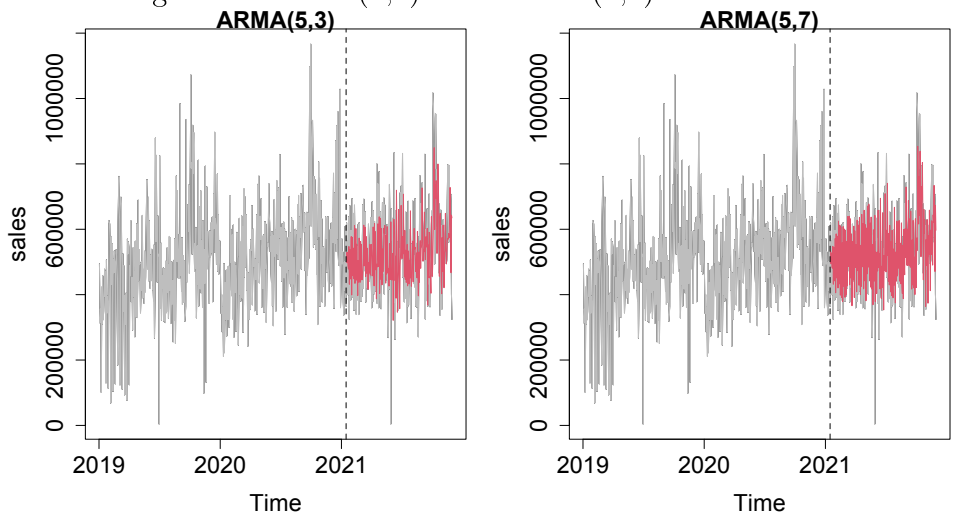
We fitted four different ARMA models with the same value for p and four different values for q. From PACF plot, we see a strongly significant spike at lag-5 so we chose p equals 5. **ARMA(5,0), ARMA(5,3), ARMA(5,5)** and **ARMA(5,7)** were picked out. Subsequently, we attained the AIC values from these model and carry out short term predictions from such model. RMSE is calculated in the next step as the square root of the mean difference between the predicted values and observed values in the test set. Table 3 summarises the RMSE and AIC of the models.

Table 3: RMSE and AIC of model classification

	RMSE	**AIC**
ARMA(5,0)	124452.3	14012.1
ARMA(5,1)	124279.6	14011.88
ARMA(5,3)	125450.9	13973.29
ARMA(5,7)	114953.6	13940.01

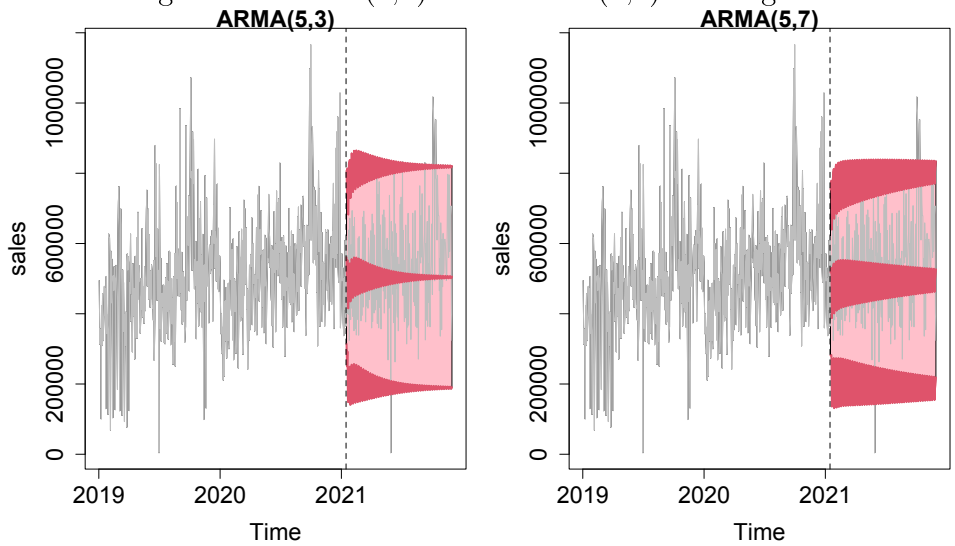
From the result of table 3, we have decided to adopt model ARMA(5,7) as best performing models to predict the values of future sales as they generate the smallest values for AIC and RMSE. We choose model ARMA(5,3) as second best performing model because of its AIC value regardless its RMSE is slightly bigger than RMSE of ARMA(5,1). The next step is to perform forecasting process with model ARMA(5,3) and ARMA(5,7). Figure 8 below shows the short term forecast plot from ARMA(5,3) and ARMA(5,7).

Figure 8: ARMA(5,3) and ARMA(5,7) for short-term



We can observe from the short term forecast plots that ARMA(5,7) predicts closer to the true values than ARMA(5,3) specially at the beginning of test set. The prediction interval of ARMA(5,7) is larger than that of ARMA(5,3). However; both models both perform adequately in generally and there is no tremendous difference between two models. Both of them can be employed to produce reliable forecast of future sale values, however ARMA(5,7) will guarantees more accuracy. The last step of our project is to generate long term forecast using the same models that have been employed for short term forecast.

Figure 9: ARMA(5,3) and ARMA(5,7) for long-term



Because ARMA(5,7) accounts for higher variance, it has somewhat larger intervals than ARMA(5,3). In long term forecast, the prediction gradually becomes a flat line in ARMA(5,3) model. ARMA(5,7) also behaves in the same way but the prediction line flattened with slower speed.

Another way to evaluate performance of ARIMA model without using metrics like

AIC or RMSE is to check the residual ACF plot and residual distribution plot. Note that in figure 10, the ACF plot of ARMA(5,3) has only one significant spike and that of ARMA(5,7) has one weakly significant spike out of almost thirty lags, which means that there is no autocorrelation left in the residuals. The residuals of ARMA(5,7) is more normally distributed than the residuals of ARMA(5,3). In conclusion, ARMA(5,7) produces the best result in forecasting future sales value but it would not lead to a significant difference in the prediction if ARMA(5,3) is opted over ARMA(5,7).

Figure 10: Residuals from ARMA(5,3)

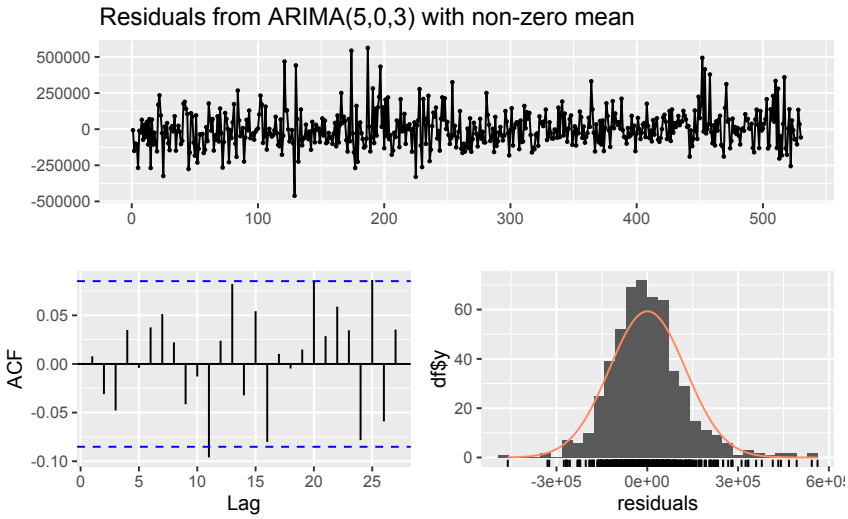
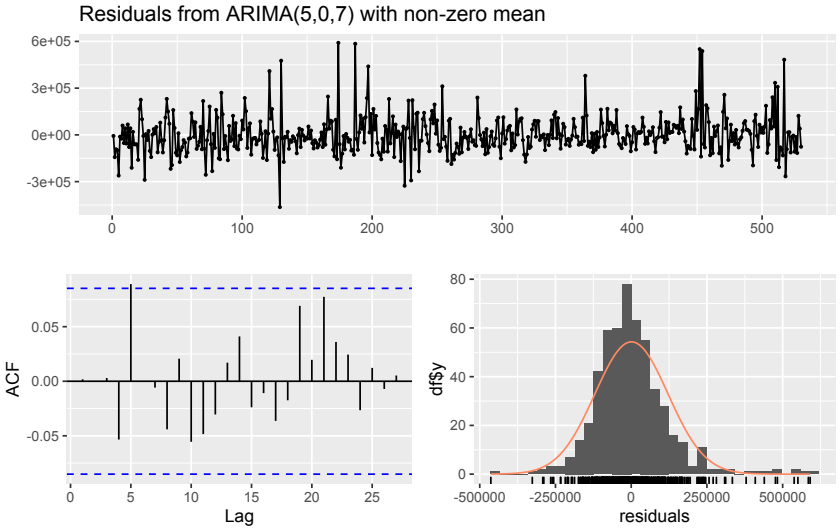


Figure 11: Residuals from ARMA(5,7)



Nevertheless, for a more accurate and better forecast, it would probably be smarter to target a company with more data available such as sales data in the weekends. This would help to construct more accurate models and better projections. However; the outcome of this analysis is still good enough to meet the demand of our clients.