Revenue Forecast Using Business Chat Text Data

Sai Vivek Kammari (A1807677)
Applied Machine Learning (7416)
Master of Data Science
University Of Adelaide

Abstract

The Project was taken up to build a pipeline for assisting small scale businesses which generate enormous text data in optimizing their performance. The primary objectives of the project are to a) generate the sales report from the financial chat and forecast the revenues for next 12 weeks considering market relevant factors such as COVID infections, Vaccination etc. We have employed various text mining techniques to generate the sales report from the text data and used boosting algorithms to forecast the revenues. The best results we could achieve are R Squared of 0.82 using an averaged model.

1. Introduction

Our project aims at mining the real time business chat text data to forecast the revenues of the business for the next three months. We aim to take various market relevant factors into consideration such as COVID- 19 infections, Vaccination rates and GDP growth rates and their impact on the business. For attaining the objectives of the project, we have explored some of the most relevant machine learning methods. For attaining the first objective that is mining the business chat text data, we have employed various text processing methodologies and libraries and for the purpose of forecasting the business revenues, we have experimented two methods broadly a) forecasting using auto regressive mythologies such as Vector Auto Regression (VAR) and SARIMAX b) the second approach is by considering the problem as a classic regression problem and make predictions using boosting algorithms such as extreme gradient boosting (XGB) and light gradient boosting (LGB). We have tried out both methodologies and compared the results obtained from both methodologies using certain evaluation metrics which are discussed in the later sections of the paper. Post comparing the results we will have finalized the methodology and concluded the project.

2. Motivation

Every business plans to better their performance by analyzing the data generated within the system. However, small businesses cannot afford to engage external consultants to analyze their business scenario and make predictions which can help them to plan their activities such as staffing, marketing budgets, expenses etc. Small businesses generate huge data in terms of business chat but do not have a tool inject this data and produce analytical reports. In some cases, businesses do use some software but that might not serve their purpose as expected.

3. Previous Research

Hitoshi Iwasaki and Ying Chen[1] attempted to use the analyst reports text data to mine and extract features for predicting the stock prices. The study also used sentiment analysis techniques though text mining to analyze the impact on the stock prices. The study employed deep neural networks to extract the text topics, the study also noticed significant improvement in the predictions and model performance by including the topic tone. Information extraction techniques such as relation extraction and event extraction were also explored by various studies for forecasting the future stock prices. The study conducted by Xiao Ding et al. [2] made use of the news articles text data and extracted events from it, in the next step the study analyzed the short term and long term impacts of the events on the stock prices using a deep convolutional neural network. The study has compared the S&P 500 index predictions and individual stock price predictions with other state of art algorithms and noticed a significant improvement in the predictions by analyzing the news articles.

Some of the studies even attempted to analyze real time streaming data and their impact on the stock prices. Sushree Das et al.[3]used the streaming data from twitter along with the stock prices data as data source and carried on a classification task using deep learning. The text data was used for analyzing the sentiment and accordingly adjust the forecast in the stock prices. A similar approach was proposed by Jiahong Li,et. al.[4] where the researchers used the forum posts to analyze the sentiment Naïve Bayes and finally performed a classification using LSTM. The

study stated that they could attain better predictions than the contemporary studies by analyzing the investors sentiment using text mining techniques.

4. Novelty of The Methodology

The methodology we presented is novel compared to previous research in this direction in terms of generating the time series data. Majority of the studies we have come across till now tried to analyze the sentiment or in broader terms the impact of other features presented in the text data on a time series data[12][13]. However, in this project, we have raw business chat text data from a real time business which the business intends to use for optimizing their operations and gain better understanding of their local market. The data preprocessing and generating the time series data for the daily sales and weekly sales from the text messages was carried out using various text preprocessing methods such as lemmatization, building the new stop words lists identifying nouns and foreign language (other than English) from the text. The other unique aspect about the project is to try to analyze the direct impact of contemporary market factors such as COVID-19, Vaccination on small businesses.

5. Methodology

Methodology: This section primarily consists of three parts namely data collection and preprocessing, proposed machine learning method and alternative methods planned to explore.

5.1. Data Collection and Preprocessing:

The data used for the project is business chat text data which focused on the financial transactions of the business. The data is fetched from a small-scale real-time business. The business chat text data has a total of 9414 text messages starting from August 2017 to July 2021. From initial exploratory data analysis, we could understand the distribution of text messages on the timeline axis and the contribution of each author in the text. We have also analyzed various plots to understand the distribution of text messages over the weekdays and over different hours of the day. The figure 1 shows the distribution of text messages with respective authors over the years, this gave some deep insights such as the most active members in the chat and the most active period of the chat. Figure 2 shows the distribution of text messages over the years.

The problem we are handling is basically a multivariate time series forecasting problem however, we have the data in the form of text. So, we had to preprocess the text data using regular expressions and nltk library in python to generate the sales report first. Some of the major steps followed in the data preparation are tokenization, removing stop words, removing special symbols, handling emojis and emoticons, handling multiline text messages, classifying the messages into sales message, expense message, general message categories, extracting the numbers from the text messages. Some of the most important preprocessing steps followed are explained below.

Handling emoticons and emojis: we have used the emot library in python to handle the emojis and emoticons. We have written a separate function to convert the emojis and emoticons to text.
Handling special symbols: We have built a list of text special symbols by scanning the input data and written a function to filter the data from these special symbols.
Removing the Stopwords: we have used the English stop words from the nltk library and built an hinglish stop words list consisting of the local language stop words and some nouns specific to the data.
Extracting the sales number from text: Also, we have defined a function to extract the numeric from the text data. The numeric extracted can be the sales number or a phone number or a date or some random number in the later stages of preprocessing we will classify identify the sales number and extract it to generate the sales report.
Handling multiline data: We have decided to split the multiline text data and create a separate record for each line in the data frame. Though, this has significantly increased the data size, it will be helpful in obtaining the objectives of the project.
Tokenization: We have used the nltk library and the lemmatization process in this step. Tokenization helped us to identify the mode of transaction and define an exhaustive list of words corresponding to sales.
Filtering the data frame: We have manually scanned the messages to identify the pattern of messages corresponding to sales. Using these patterns we have filtered the data frame to obtain the messages corresponding to sales
Extracting the mode of payment: In the data for every payment done, the mode of payment is mentioned, we have made an exhaustive list of payment modes from the data and defined a function which outputs the payment mode.

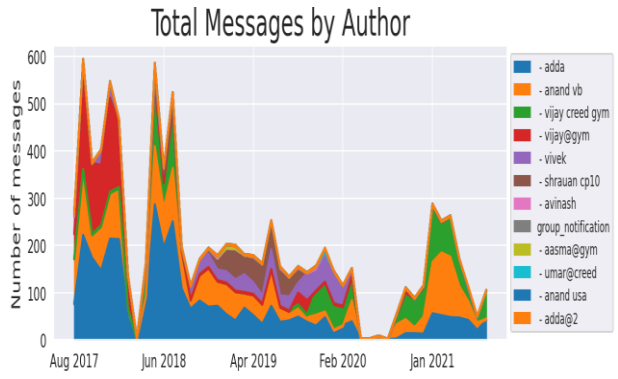

Figure 1: Distribution of text messages by authors

In the final step of the preprocessing, we have obtained the sales report, we have noticed that due to business operations hurdles or any unknown factors, sales on some of the days were zero and we have also noticed that the distribution of zero sales days is very random.

Market relevant data factors such as COVID-19 infection rates, hospitalization rates, vaccinations and GDP data were manually collected from Institute for Health Metrics and Evaluation[5] and Trading Economies respectively[6]. The final preprocessed data has weekly sales, COVID 19 infections, Vaccination percentages, number of lockdown days in the week and GDP growth rate for the quarter. Now we can use suitable multivariate time series forecasting algorithms to generate the forecast.

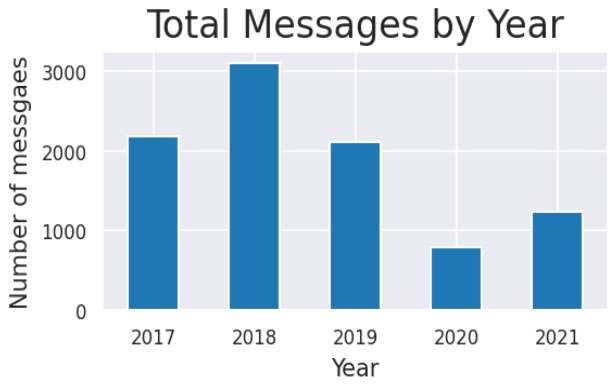

Figure 2: Distribution of text messages across years

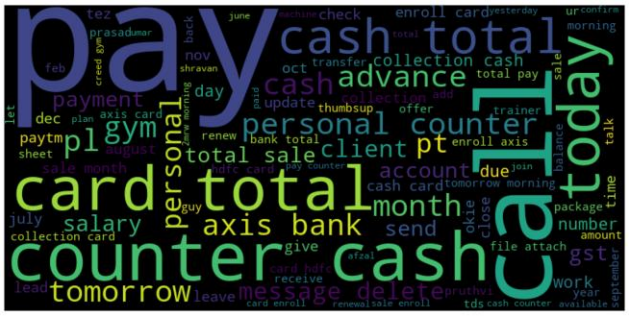

Figure 3: Word cloud highlighting the main subjects of the text data.

5.2. Machine Learning Algorithms

For forecasting the business revenues, we have experimented two methods broadly a) forecasting using auto regressive mythologies such as Vector Auto Regression (VAR) and SARIMAX b) the second approach is by considering the problem as a classic regression problem and make predictions using boosting algorithms such as extreme gradient boosting (XGB) and light gradient boosting (LGB).

Auto Regression Methods: We have explored various time series forecasting algorithms and studied the best suited scenarios for all the algorithms. After conducting a thorough research, we have finalized the Vector Auto Regression (VAR) algorithm for predicting the future sales of the business. Vector auto regression is the extension of autoregression algorithm and is primarily used where we have multiple timeseries inputs and there exists a complex relationship between these independent time series [7]. In the vector auto regression model each variable is assumed to be a linear function of its past values and the past values of other variables.

$$y_1(t) = a_1 + w_1 11 * y_1(t-1) + w_1 12 * y_2(t-1) + e_1(\text{t-1})$$

$$y_2(t) = a_2 + w_1 21 * y_1(t-1) + w_1 22 * y_2(t-1) + e_2(\text{t-1})$$

$y1$,$y2$ are time series variables with t and t-1 representing the current value and first lag values. a1 and a2 are the constant terms. w11, w12, w21, and w22 are the coefficients. e1 and e2 are the error terms.
A first order vector auto regression can be represented by the above equations. It is clear from the equation that the value of a variable in dependent on one past value of itself and one past value of other variable as we are just considering the first lag, it is called as a first order vector auto regression[1].The method is suitable only for multivariate time series without trend and seasonal components. To make sure that the data is fits the algorithm we will have to check following properties of the data.

The vector form representation of the above two equations representing the first order vector auto regression can be seen in the above equation. The method is suitable only for multivariate time series without trend and seasonal components. To make sure that the data is fits the algorithm we will have to check following properties of the data.

Stationarity: If the statistical properties such as mean and variance of the timeseries data remain constant or have relatively small variation over the time then the timeseries is said to be stationary [8].

Causality: The basic criteria of the vector autoregression model is that each time series has an impact on the other timeseries. This can be checked using the Granger's Causality Test.
Boosting Algorithms: We have used some of the conventional machine learning predicting algorithms such as XG Boost regressor, and Light Gradient Boosting to predict the future sales. All the models we have explored

belong to the class of boosting algorithms. Boosting is a technique in which a certain weak learning algorithm is scaled up to make the training error zero. In general, boosting is not a single algorithm it is a family of algorithms.

We have used two models on each of the datasets during the experiments and noted the results from both the boosting models were significant. Furthermore, we have strived to improve the predictions by combining these two algorithms. For combining the three boosting algorithms we have followed two methods.

Averaged Model: In this methodology we have taken the three boosting models and trained them on the same train data and made three different predictions using the two boosting algorithms. The final predictions were calculated by simply taking the average of the two predictions.

Final Predictions = Average (XGB predictions, LGB predictions)

We have noticed that the averaged model outperformed all the three boosting algorithms in all the experiments.

Meta Learner: In this methodology, similar to the averaged model, we have trained the two boosting algorithms on the same train data and made two different predictions using the two algorithms. The two predictions and the actual output were put together to form new data and a LASSO regressor was trained on this data and the final predictions made by the LASSO are taken as the final predictions.

Final Predictions = LASSO (XGB predictions, LGB predictions)

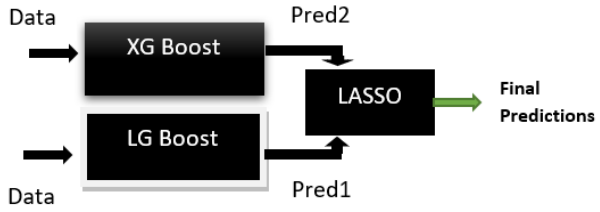

Figure 4: Stacked Regression of XGB and LGB using LASSO Meta learner

5.3. Evaluation Metrics

Evaluation metrics form the core part of the machine learning project as they help us to measure the performance of the machine learning methodology we have employed. In our case, the project on a broader level is a regression problem so we cannot measure the accuracy of the methodology however, we have special class of metrics to measure the performance of the regression methodology. We have chosen Root Mean Squared Error (RMSE) and R-Squared.

R-Squared: R-Squared is a metric to evaluate how good the employed regression model has fit our data. The main principle is to compare the residual sum of squares denoted as SS(res) and the total sum of squares denoted by SS(total). SS(total) is the sum of square of distance of the points from the average regression line. SS(res) is the sum of the square of distances of all the points from the best fitted regression line [9]. The value of R-Squared generally ranges from 0 to 1. The closer the value of R-Squared to 1 the better the regression model. In some cases, the R-squared value is also negative, and this negative value represents that the regression model has performed worse than the average model.

One of the major drawbacks of R-squared is that, whenever a new attribute is added to the regression model, the R-Squared value always increases without checking the significance of the attribute.

$$(SS_{total}) = sum(y_i - y_{avg})^2$$

$$(SS_{res}) = sum(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - (SS_{res})/(SS_{total})$$

Root Mean Squared Error (RMSE): From the above formula we can state that RMSE is the square root of the average squared error. It measures the difference between the actual value and the predicted value and is the measure of how well the regression line fits the data points.

6. Results & Discussion

Apart from forecasting the future revenues, one of the prime objectives of the research was to mine the text data and generate the sales report. Prior to entering the second phase of forecasting the revenues, we made sure that the generated sales report matches with the monthly and annual sales numbers shared by the business with us. The generated sales report was compared with the actual monthly and annual sales numbers. We have noticed that the generated sales report was an exact match with the actual sales report.
need to make sure that we are using the correct data.

Autocorrelation Check: As discussed in the earlier sections and the methodology section of this report we have initially thought that the problem we are addressing can be solved by using multivariate time series auto regressive methods. However, the autoregressive models assume the timeseries observations at previous time steps are useful to predict the value at the next time step. For using the auto regressive methods, we have conducted some statistical tests to check

the auto correlation of the sales data. From the lag plots, autocorrelation plots and the calculated correlation between the timeseries and its lags we could notice that autocorrelation was not very significant and hence our idea of using the auto regressive methods ended futile.

predict the future values of the other time series.
For using VARMAX or SARIMAX the cause-and-effect relation should exist in two directions i.e. a variable x should cause y also y should cause x. From the results of this test, we could observe that the timeseries sales data was

Figure 5: Weekly Sales prediction using Averaged Model.

Though we have observed the autocorrelation in the sales data is not significant, we have moved ahead and conducted the Augmented Dickey–Fuller test to check the stationarity of the timeseries data and the Granger causality Test [10.

Augmented Dickey–Fuller test: Stationarity of the data can be tested by visually inspecting the data for any trend or seasonal occurrences or by performing the unit root test. The presence of unit root in the time series makes the time series non-stationary. The Augmented Dickey–Fuller test Is the extension of Dickey–Fuller test to include higher order lags [11]. From the results of this test, we could observe that the timeseries data was stationary and can be modelled as a timeseries.

$$y\ (t) = c + \beta(t) + \propto * y\ (t-1) + \emptyset \Delta Y(t-1) + e(t) \quad ----(1)$$

$$y\ (t) = c + \beta(t) + \propto * y\ (t-1) + \emptyset_1 \Delta Y(t-1) + \emptyset_2 \Delta Y(t-2) + \cdots . + \emptyset_p \Delta Y(t-p) + e(t) \quad ---(2)$$

The above two equations represent the Dickey–Fuller test and The Augmented Dickey–Fuller test respectively.
$y\ (t-1)$ is the first lag of time series and ø(Δ) Y(t-1) is first difference of time series at time(t-1).

Granger causality Test [4]: Generally, granger causality test is performed to check the correlation or existence of cause-and-effect relation between independent timeseries variables. In other words, one time series is useful to

caused by COVID timeseries however, the vice versa was proved to be wrong.

From the summary of the autocorrelation check, Augmented Dickey–Fuller test and the Granger causality Test we have decided the timeseries problem is not a good fit for the initially thought methodology and hence decided to take the other route of solving it as a regression problem as we knew there is significant correlation between the COVID factors and the sales of the business.

Boosting Algorithms: Post concluding that the autoregressive methods will not yield better results, we have performed featuring engineering techniques to capture market relevant features such as Lockdown days, COVID vaccinations, mobility, footfall, and deaths. To form the final dataset we have combined these features with the sales data.

Table 1: Model performance comparison for Weekly Sales

Model	R Squared	RMSE
XGB	0.7804	14771
LGB	0.7260	16498
Stacked Regression	0.7089	17007
Averaged Model	**0.8202**	**13366**

Figure 6: Monthly Sales prediction using Averaged Model.

By analyzing the correlation using heatmap plot, we could notice that the sales data is significantly correlated to the market relevant features. On this final dataset, we have used XG Boost and LG Boost algorithms to make the predictions.

For predicting the future revenues, we have taken up three approaches, i.e., predicting daily sales, predicting weekly sales, and predicting the monthly sales.

Daily Sales: In an attempt to predict the daily sales, we have used the generated sales report in the preprocessing module. The main challenge in this attempt was, on many days the sales were zero and we have also tried to analyze the distribution of zero sales days but, we could not come up with significant trend or rather we could notice they were very random. We strongly believe the zero sales days are more correlated to factors pertaining to business operations such as staffing, business operating hours, local holidays etc. As on the date of building the project, we did not have data to model the business operations and hence we have not gone ahead to predict the daily sales of the business.

Weekly Sales: To nullify the impact of the effect of business operational features such as staffing, operating hours on a given day and local holidays, we have decided to calculate the weekly information of all the features and sales. We could notice that the target variable still had large variance, we understand that is because of various market factors such as the lockdown days, COVID infection rates, Vaccination numbers, Deaths, and hospitalizations. The figure 6 shows the performance of various machine learning models on the weekly data and we could notice that the Averaged model of XG Boost and LG Boost, where Final Predictions = Average (XGB predictions, LGB predictions) are best. In terms of the evaluation metrics the Monthly Sales: We have also made an attempt to generate a forecast of the monthly sales. The reason backing this attempt was, rather than the daily sales and weekly sales we could notice that the monthly sales had a clear trend. The plot in figure 6 shows the prediction of the monthly sales data using the averaged model which has given the best performance with the weekly data. In terms of evaluation metrics, the R squared value is 0.8917179 and RMSE is 44265. Though we cannot directly compare the weekly sales predictions and monthly sales directly but, by looking the relative evaluation metrics such as R squared we could argue that the monthly predictions are relatively better than the weekly sales predictions and which in turn are better than the daily sales predictions.

7. Learning Outcomes

Through the research done during the project, I have explored and learned various time series forecasting methods and got a deeper insight of the timeseries properties such as auto autocorrelation, stationarity, causality. To check the properties of the timeseries, I have also gained knowledge of various statistical tests such as adfuller test, granger causality test and their interpretations. The other major challenge in the project was to mine the unstructured text data, as this was my first project working on the text data, I have learned various text mining techniques, I have also gained significant knowledge and practical understanding of basic NLP concepts such as lemmatization, tokenization, stop words, parts of speech tagging, relation extraction, named entity extraction. In the process of exploring the previous research works, I have gone through various deep learning methodologies used in the text mining feature extraction and stock price prediction, these concepts have dragged my attention and pushed me to make build a future plan for learning and diving deep into NLP.

8. Future Work

As discussed in the earlier sections of the report, though we have achieved significantly good predictions on the monthly data. We strongly believe that the weekly sales predictions can be further be improved. On the other hand, we could not model the daily sales as the daily sales are more dependent on the business operational features such as staffing, business operating hours and local holidays. In the future we would like to come up with a methodology to capture these features and their correlation with the daily sales. We believe by including features pertaining to day-to-day business operations, we can model the daily sales and improve the weekly sales predictions. Finally we also plan to work on the deployment part of the model and develop an user interface which can take in the business chat text input and outputs the sales report and give predictions for the next three months or 12 weeks.

9. Conclusion

Through the research we could obtain the set objectives of the project, that are generating the sales report from the business chat text data and forecasting the business revenues with good confidence. The objectives were achieved by employing various text mining techniques and machine learning techniques. In the various sections of the report we have discussed the detailed implementation and procedures followed and presented the results.

References

[1] Hitoshi Iwasaki and Ying Chen. Topic sentiment asset pricing with dnn supervised learning. SSRN Electronic Journal, 2018

[2] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, pages 2327– 2333. AAAI Press, 2015.

[3] Sushree Das, Ranjan Kumar Behera, Mukesh Kumar, and Santanu Kumar Rath. Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia Computer Science, 132:956–964, 2018

[4] Jiahong Li, Hui Bu, and Junjie Wu. Sentiment-aware stock market prediction: A deep learning method. In 2017 International Conference on Service Systems and Service Management. IEEE, June 2017.

[5] Institute for Health Metrics and Evaluation COVID-19 (healthdata.org)

[6] TRADING ECONOMICS India - Economic Forecasts - 2021-2022 Outlook (tradingeconomics.com)

[7] Ericsson, NR & Reisman, EL 2012, 'Evaluating a Global Vector Autoregression for Forecasting', International Advances in Economic Research, vol. 18, no. 3, pp. 247– 258

[8] Shahin, MA, Ali, MA & Ali, ABMS 2014, 'Vector Autoregression (VAR) Modeling and Forecasting of Temperature, Humidity, and Cloud Coverage', in Computational Intelligence Techniques in Earth and Environmental Sciences, Springer Netherlands, Dordrecht, pp. 29–51.

[9] Stephanie Glen. "Coefficient of Determination (R Squared): Definition, Calculation" From StatisticsHowTo.com: Elementary Statistics for the rest of us!
https://www.statisticshowto.com/probability-and[1]statistics/coefficient-of-determination-r-squar

[10] Su, C, Xu, Y, Chang, HL, Lobont, O-R & Liu, Z 2020, 'Dynamic Causalities between Defense Expenditure and Economic Growth in China: Evidence from Rolling Granger Causality Test', Defence and Peace Economics, vol. 31, no. 5, pp. 565–582.

[11] Hamilton, JD 1994, Time series analysis , Princeton University Press, Princeton, New Jersey.

[12] 홍성혁 & Hong, S 2020, 'A study on stock price prediction system based on text mining method using LSTM and stock market news', 디지털융복합연구, vol. 18, no. 7, pp. 223–228.

[13] Obst, D, Ghattas, B, Claudel, S, Cugliari, J, Goude, Y & Oppenheim, G 2019, 'Textual Data for Time Series Forecasting'.