Capstone Write Up

Name of the Student

Md Tajrianul Islam

islam.mdta@northeastern.edu

ALY6980 23293 –Experiential Capstone

Summer 2021 Quarter

Submitted to –Eugenia Bastos

MPS Analytics, College of Professional Studies, Northeastern University

Introduction

Our venture support Eriyam Inc has fostered an item dependent on PathosAI, which alludes to enthusiastic commitment as a basic factor to comprehend the excellent powerhouses of a buy choice. It is significantly more compelling than simply investigating surveys with the assistance of NLP to comprehend whether a purchaser has a fortunate or unfortunate involvement in the item. It's normal said about audits and overviews that a great many people can't communicate our intentions/sentiments behind a specific buy choice. So frequently the outcomes we get from the overviews or surveys can't actually assist a business with rolling out the important improvements. As per crafted by Coppola (2021), in 2019 almost 1.92 billion clients have purchased items and additional benefits on the web. In 2020, in spite of the pandemic's adverse effects on individuals' buying designs, online business deals volume became by practically 28% and established $4.28 trillion around the world (Cramer-Flood 2021). In this age and time it turns out to be more imperative to use this information and discover significant data that can help in the business interaction.

Our sponsor has given us a dataset that contains 15,365 columns of surveys of child items from various drug organizations, determined enthusiastic commitment with PathosAI, and if the shopper will suggest the item. The dataset additionally holds data about the date the surveys were made, what was the driver behind the deal with Ad reviews, when tone and taste of the items were referenced in the audits, and various phases of their client venture. We will probably investigate the current information alongside outer data about the organizations so organizations can get noteworthy data from our constructed models that can assist them with expanding their deals.

Research Hypothesis

Our research process is based on the hypothesis that, being able to understand the factors that influence emotional engagement will enable actionable information for businesses to act upon to maximize their sales.

We want to propose and build a model that can tell us which attributes affect the emotional engagement value the most. Our goal is to find out the significant attributes in the prediction model, where our target variable will be emotional engagement.

Methodology

We have planned to try out models like multiple linear regression, support vector machine and decision tree. All of them are supervised learning. So we can guess that the data preprocessing will include a hand-coded small set of documents for whatever variable(s) we care about depending on the EDA. Then we will apply the models and finally compare the results and specify the important attributes.

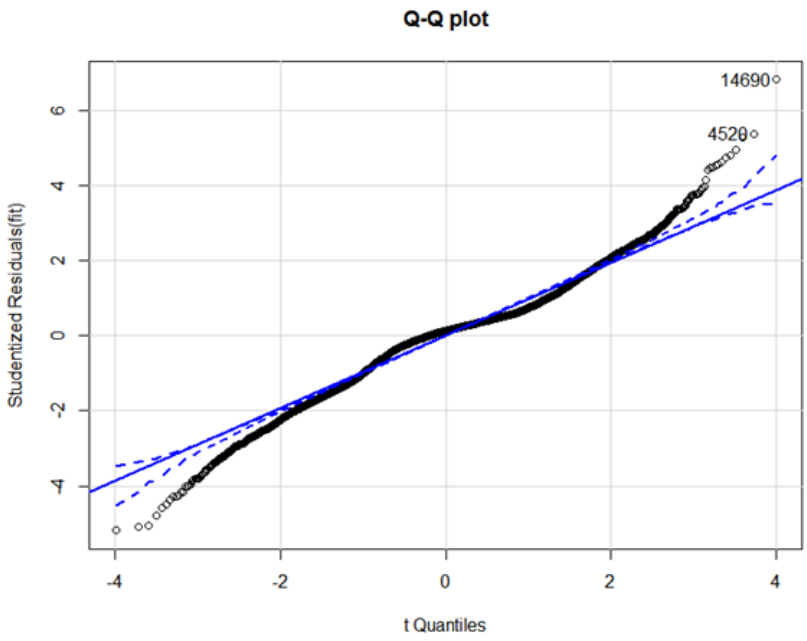
Literature Review

I assessed 5 diverse friends and checked on articles that covered spaces of how to develop an examination speculation for shopper conduct investigation to what kind of calculations can be utilized for such purposes. I initially began with looking into shrewd retail innovations, which recommends three unique factors that impact client commitment: conduct execution anticipation, exertion hope, impacts (social and social aim on genuine conduct). Social impact is a major factor also which makes them trust the frameworks to totally finish their work. The article likewise thinks about the impact of social expectation on real conduct. For the best outcomes the creators have included segment factors, for example, age gathering, and sexual orientation as control factors to take into account better depiction of the relationship proposed in our model and to give a more thorough trial of the hypothetical linkages. Then, at that point I likewise took a gander at how we can make prescient models to break down driving conduct. As it manages comparative printed information, we discovered that of all the picked research that took on NN for DB examination, about 57% used DNNs through 43% used conventional ANNs. The article furthermore proposes standard execution appraisals for ML models. I likewise saw some buy conduct demonstrating accomplished for internet business sites for our exploration references. In this paper, they study brand buy forecasts by exploring rehearsals, which might provoke brand purchases. They mention three observable facts. (1) they explored veritable electronic business data from various focuses. Focusing in on customers' picture purchases, they sort out practices' advancement with time and practices' collaboration. (2) For different practices, they eliminate

assorted time-creating components that can fill in as pointers of customers' picture purchase. (3) They use an essential backslide based model by changing the limits of time creating components and others in two exceptional circumstances (the progression buy estimate and the regular buy number assumption) to assemble two tests. The examination results show that the model using three kinds of features plays out the most incredible in the two circumstances, and the time-creating component expects the fundamental part among them.

Data Analysis

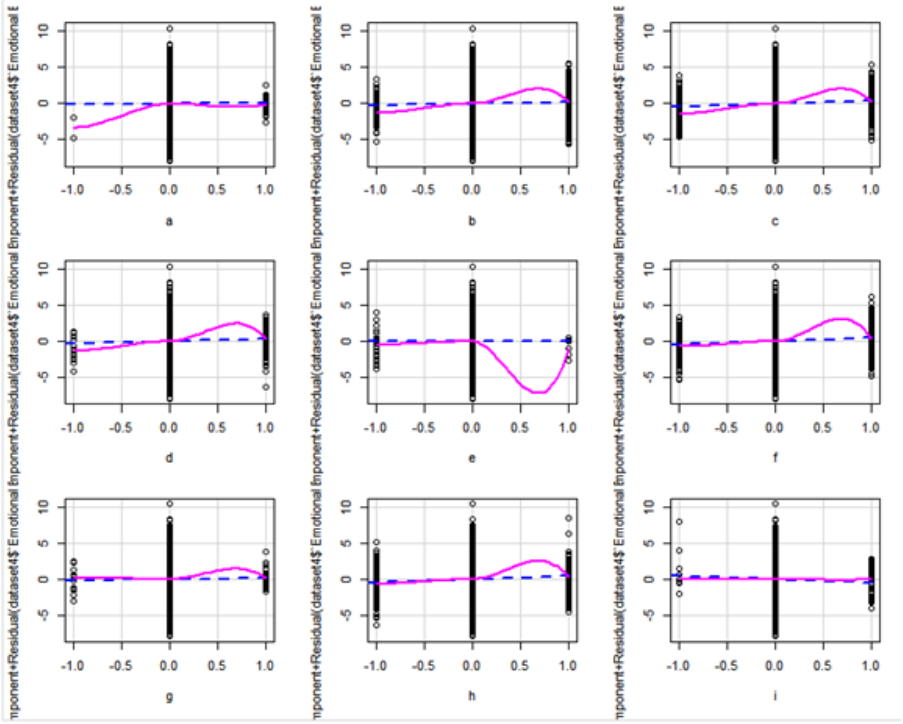
To start the analysis, I want to transform the data and make it ready to run the model. Although, for the final project I did it in a different way, where I manually encoded some specific columns for them to be used in supervised learning. To prepare the data for that, I uploaded the data, checked for spelling mistakes, null values were identified and removed, and finally data types for each of the columns were formatted. The Emotion column was divided into three categories: positive (containing one or two positive emotions), negative (containing one or two negative emotions), and unclear (containing both positive or negative). Values were classified based on that. Next, Emotional Engagement was transformed into a numerical variable. And then worked on the models. After making the multiple linear model, a hypothesis test was performed. Regression analysis needs to meet some preconditions. First, Q-Q plot was used to check normality. We can see that all the points are near the line, while some of them are not in the confidence interval. We can conclude that the data conforms to the normality hypothesis.

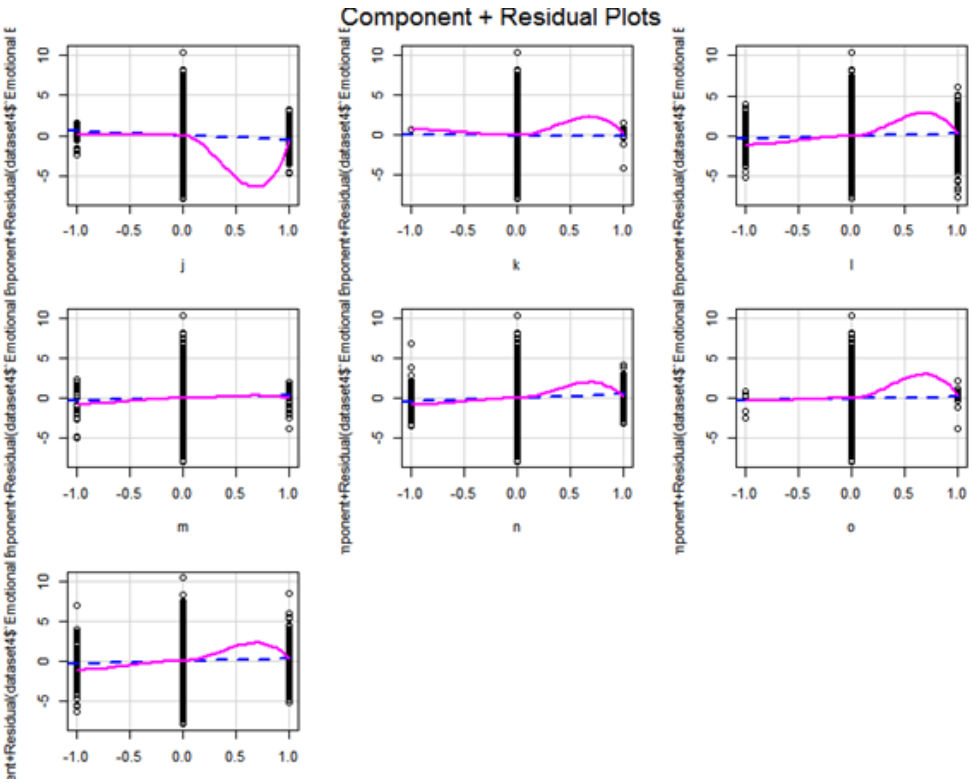


We then use the D-W test to test for independence. It can be seen that P=0.126>0.05, indicating that there is no autocorrelation between variables and they are independent.

```
> durbinWatsonTest(fit)
 lag Autocorrelation D-W Statistic p-value
   1      0.01228086      1.975429   0.126
 Alternative hypothesis: rho != 0
```

Then we check the residual graph, and it can be seen that the component residual graph confirms the linear hypothesis, indicating that the linear model is more suitable for this data set.

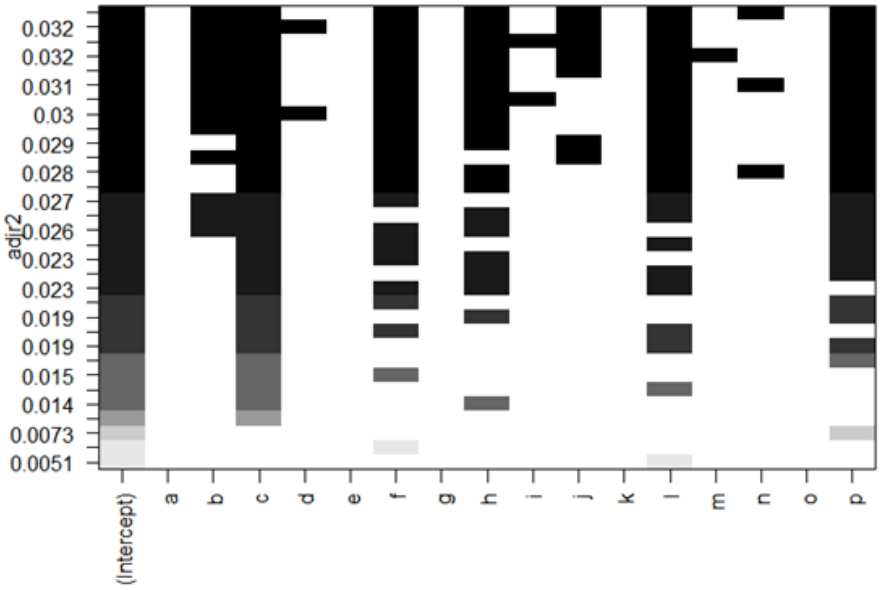

a
b
c
d
e
f
g
h
i


Component + Residual Plots
j
k
l
m
n
o

And then we check for homoscedasticity. The result of p<2.22e-16 is significant, indicating there is heteroscedasticity.

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 167.1236, Df = 1, p = < 2.22e-16
```

Then in order to find the best model, leaps() function was used for a full subset regression. We know that the larger the adjustment R squared is, the better the model fit degree is. Compared with the previous model, these independent variables have passed the significance test. Therefore, b,c,f,h,j,i,n, and p were chosen as independent variables to model again.



Next, an SVM model was built .Strictly speaking, it is an SVR model because the dependent

```
> data.frame( R2 = rsquare(fit1, dataset4),
+             RMSE = rmse(fit1,dataset4),
+             MAE = mae(fit1, dataset4))
          R2     RMSE      MAE
1 0.03005528 1.534814 1.117195
> data.frame( R2 = rsquare(svm.model1, dataset4),
+             RMSE = rmse(svm.model1,dataset4),
+             MAE = mae(svm.model1, dataset4))
          R2     RMSE      MAE
1 0.04550749 1.530015 1.061343
```

variable is continuous.

The R squared, RMSE and MAE of the two models were compared. The results were similar. In terms of RMSE, SVM model has smaller error, so we can use this model to predict Emotional Engagement. However, due to the principle of the SVM model, we cannot know the importance of each variable. Considering that the RMSE of linear regression is similar, we can know which independent variables are important through the parameters before each variable of linear regression.

In this model, all independent variables passed the significance test.

```
Call:
lm(formula = dataset4$`Emotional Engagement` ~ b + c + f + h +
    j + i + n + p, data = dataset4)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7801 -0.7075  0.2097  0.7692 10.4175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.37833    0.01317 104.631  < 2e-16 ***
b            0.28211    0.04223   6.680 2.46e-11 ***
c            0.46841    0.04905   9.550  < 2e-16 ***
f            0.45070    0.05927   7.604 3.04e-14 ***
h            0.48828    0.07250   6.735 1.69e-11 ***
j            0.57158    0.12172   4.696 2.68e-06 ***
i            0.45010    0.20806   2.163   0.0305 *
n            0.48651    0.10175   4.781 1.76e-06 ***
p            0.34971    0.04467   7.829 5.24e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.535 on 15353 degrees of freedom
Multiple R-squared:  0.03006,   Adjusted R-squared:  0.02955
F-statistic: 59.47 on 8 and 15353 DF,  p-value: < 2.2e-16
```

Conclusion

For the capstone project, we have applied multiple linear regression and support vector machines to predict the emotional engagement. Also as a group we concluded that SVM is a better model to predict emotional engagement. We also observed that when emotions are less than 1 (negative) , the engagement mean is around 7.4, but when larger than 1 then it depends on whether customers felt "value for their money". If they did, the average engagement value was 9.4, representing 7% of the actual data. Lastly, the following variables are the most significant: satisfaction, switch, promotion, delivery, return, refund, smell, and recommendation. We also see the parameter for refund and return are the highest, meaning the quality of refund and return will ensure whether the consumer is emotionally vested in the product or not.

References

Coppola, D. 2021. E-commerce worldwide - Statistics and Facts. Statista. Retrieved from: https://www.statista.com/topics/871/online-shopping/#:~:text=As%20internet%20access%2 0and%20adoption,3.5%20trillion%20U.S.%20dollars%20worldwide.

Dong, Y., & Jiang, W. (2018, January 23). Brand purchase prediction based on time-evolving user behaviors in e-commerce. Retrieved August 08, 2021, from https://onlinelibrary-wiley-com.ezproxy.neu.edu/doi/pdfdirect/10.1002/cpe.4882 DOI: 10.1002/cpe.4882

Igorevic, B. S. (2021). Advanced Analytics for Prediction of Customers' Preferences: L'Oréal Case. Retrieved August 08, 2021, from https://dspace.spbu.ru/bitstream/11701/30005/1/Masters_Thesis_Bruchkus_Vlasova.pdf