Title
Statistical Analysis of Customer Satisfaction Scores (NPS) in the Direct-to-Consumer Clear Aligner Industry

Author
Chang, Wilfred

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Statistical Analysis of Customer Satisfaction Scores (NPS) in the Direct-to-Consumer Clear Aligner Industry

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Wilfred Chang

2022

ABSTRACT OF THE THESIS

Statistical Analysis of Customer Satisfaction Scores (NPS) in the Direct-to-Consumer Clear Aligner Industry

by

Wilfred Chang
Master of Applied Statistics
University of California, Los Angeles, 2022
Professor Frederic R. Paik Schoenberg, Chair

This thesis is an analysis looking into consumer experience and the predictability of consumer feedback based on Net Promoter Score (NPS). The industry focused on in this document is the DTC clear aligner industry. Ideally, this research is meant to be applied in "real world settings" with the intention that companies can leverage this concept to help improve their product's own experience. In an effort to achieve that goal, we hope to answer two aspects 1) whether certain events that occur during a customer's product experience impact score and to what degree and 2) whether machine learning methods can accurately predict scores based on the certain events mentioned. The research begins by performing exploratory data analysis to understand event correlation to score, running a few models while comparing the accuracy and discussing the modelled results while highlighting the importance of the research for companies in the industry.

The thesis of Wilfred Chang is approved.

Guani Wu

Yingnian Wu

Maryam Mahtash Esfandiari

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Motivation

In today's day and age, consumers are hyper-sensitive about understanding what products they are buying before making the actual transaction. This often means reading reviews online and getting product ratings from other consumers. In fact, according to a study conducted by Qualtrics [Qua22] (a reputable surveying tool for consumer feedback), over 90% of consumers read online reviews before buying a product. This staggering percentage shows how widely revered product reviews are in the digital age. Moreover, based on a study led by Mckinsey [Wis] (premier Global Management Consulting Firm), growth in product reviews has grown by an average of 87% between December '19 and December '20. This is particularly true with Arts & Entertainment and Food/Beverages/Tobacco products which top the list in percentage growth change year over year. These statistics illustrates the verticality in growth of product reviews and the reliance consumers have on them in their purchasing decisions.

With reviews becoming such an instrumental aspect in purchasing decisions, companies must heed the feedback of their consumers and improve upon or craft new products around it. This rings especially true in the direct to consumer (DTC) market, where there is no middle-men helping companies sell their product on their behalf. In a recent statistic posted by E-marketer, US DTC E-commerce Sales are anticipated to top $175 billion dollars by 2023 and the number has been achieving double digit growth year on year [Gol21]. All of this is

to say that, with the growth of DTC products coupled with the consumer propensity to look for consumer product feedback, the desire for companies to stay on top of their customer satisfaction is crucial.

Reiterating the importance on customer satisfaction, it often begs the question of whether companies can use data based on their customer's experience paired with statistical methods to help predict customer satisfaction score. This thesis is meant to address that question. In this research topic, we will be analyzing company data related to the customers' experience within the transparent dental aligner industry, an industry that is growing within the DTC market [Par21]. Our response variable we will look into understanding and predicting is an industry recognized metric known as Net Promoter Score (NPS) and our goal is to understand what components within the customer journey shape the score and to what level of accuracy can we predict the score.

The motivation behind applying this topic to the clear aligner industry is that the clear aligner industry is relatively new. Being able to identify specific parameters that impact the scores could give companies an opportunity to learn and identify what factors are important in receiving a positive score. In turn, it would help the company and industry grow as they work to improve the product and experience. Moreover, being able to predict the score based on those sets of parameters could give companies advanced notice of an impending negative score and interject and work to improve the customer's experience while mitigate the chances of a lower NPS score that ultimately reflects the product and company.

1.2 Understanding of the Business

Before diving into the data, it is best to provide some background information about the industry we are analyzing about to better understand the conclusions we draw. Transparent dental aligners are essentially what it sounds like; they are aligners that attach on a customer's teeth with the hopes of straightening their smile. Similar to braces, dental aligners

generally last an extended period of time (6-18 months depending on severity) but they carry an added benefit of being removable for eating, drinking and generally whenever inconvenient to wear them. According to research by Dentistry Today [Par21], the clear aligners market is expected to grow approximately 16% annually and top $8.7 billion in market size by 2028. While convenient to wear, the steps in obtaining clear aligners can be arguably lengthy.

Figure 1.1 illustrates the process in which a customer undergoes in the beginning stages of the customer experience (e.g. order/delivery/impression-making) as well as the process while in treatment. For the purposes of this research, we focus on consumer facing experience points and remove any "behind the scenes" moments as customers most likely will not see these points and it won't reflect in their perspective and score. A more detailed explanation is provided below.

Generally, the process begins by the customer ordering the impression kit which will allow them to create a molding of their teeth. After completing the molding, customers return their impressions to the dental lab in which the lab inspects the impression kit. Good impressions proceed through the process while bad impressions will require the customers to re-perform their impression in what's called a rekit. The process continues with a treatment plan indicating the length of their treatment and how the end-result would appear in a digital render. Assuming the customer approves, the dental company checks if the required dental/medical history is provided and, if not, the customer's account is in a holding pattern until completed; this status is coined as "Stipulations" or "Stips" for short. Next, a licensed dentist reviews the plan and, upon approval, the plan is sent to manufacturing to craft the aligners. After manufacturing is complete, the aligners are sent by the company to the end customer as they officially start the treatment phase. During the treatment phase, customers have the option to "check-in" which allows customers to provide images of their progress and indicate their status (overall fit/discomfort/etc). Also during treatment, customers can open up "cases" which is essentially logging an issue and loops in various customer service teams

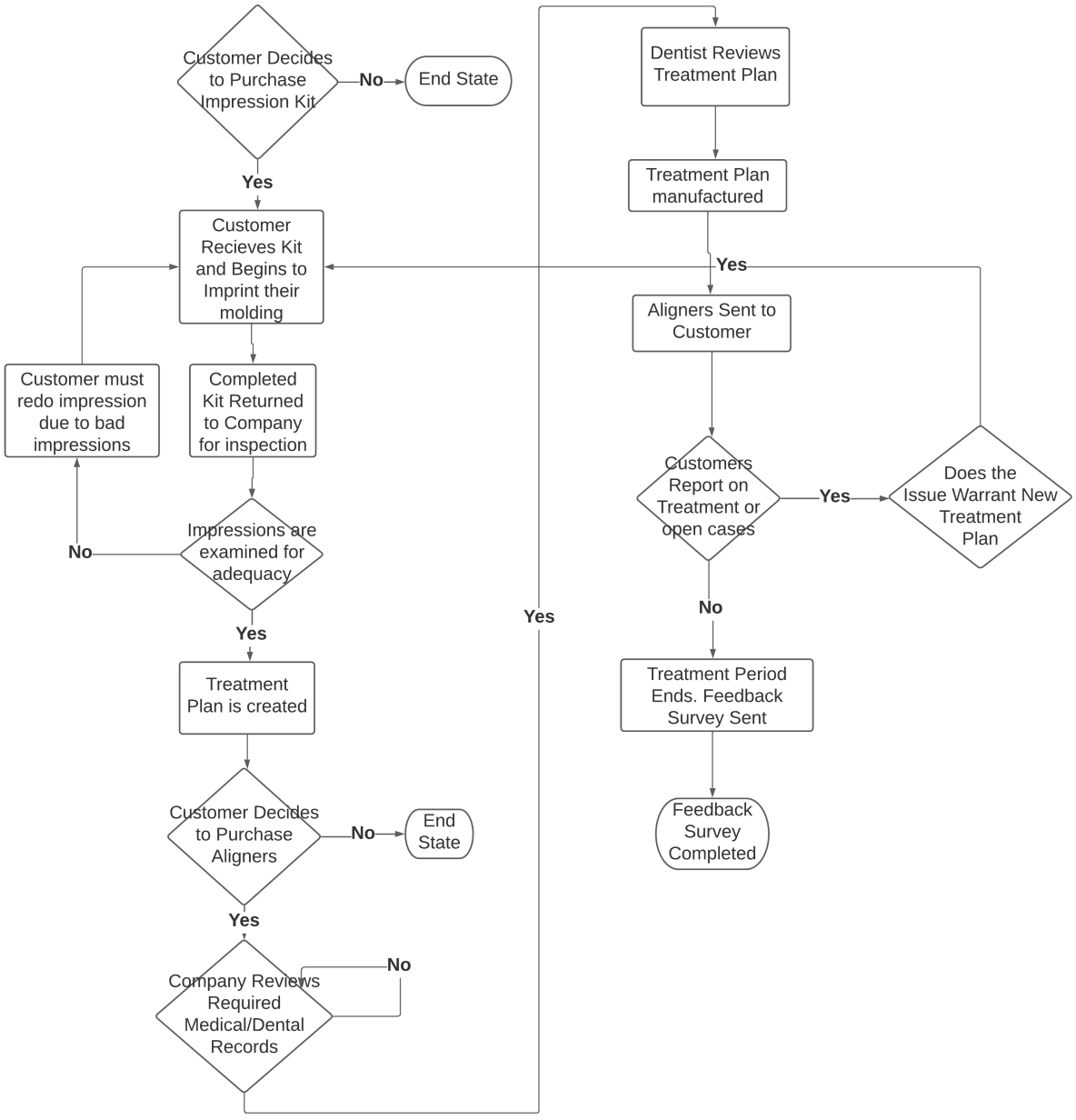

Figure 1.1: Dental Aligner Process Flow (Source: Author)

depending on the severity. In some cases, if it is found that the plan is not on track or there is too much discomfort in the customer's treatment, the customer will be required to restart the impression process in what's called "refinements". Finally, assuming no problems and no refinements are needed, customers indicate their treatment is complete and the dental

company sends out a survey asking for feedback and a score in the form of an NPS survey.

1.3 Understanding of Net Promoter Score

Net Promoter Score, NPS for short, is a metric that companies of all industry verticals use to understand how well their product is performing in the market. While the question can be framed in many different ways, the crux of the question asks about the likelihood that the customer would recommend the product to a friend or colleague. Generally, the response the user inputs is between a 0-10 scale but, for our analysis, we will focus on 3 levels derived from the score; promoters, detractor and passive (neutral). Detractors (0-6) are generally customers that found the product disappointing or something within the product experience let them down that fundamentally impacted their experience with the product. Promoters (9-10) are customers that are satisfied with the product and had a positive experience. Finally, passive (7-8) customers are generally those who liked the product but flaws impacted their experience. NPS scores are one of the only feedback loops companies receive in order to improve and iterate upon their product which illustrates their importance. In our research context, we will try to understand what factors impact NPS scores and to what level of accuracy do models carry in predicting the score within the clear aligner industry.

CHAPTER 2

Data Structure

This chapter is intended to help inform readers on the origin of where the data we are analyzing came from, an explanation of the features we look into and clarification on how the data was processed.

2.1 Origin

The data comes from a company that specializes in clear aligners in the Direct to Consumer industry. Due to privacy concerns, the company name is withheld but the records are customers who started the product experience by purchasing an impression kit, completing treatment and filling out the NPS survey post treatment. Tables 2.1-2.5, depicts the full set of features that was proposed that potentially had an impact on a customer's NPS score. The list was based on human understanding of the business and no prior analysis was done before to confirm the intuition.

Table 2.1: ***Score Response Variables*** Feedback score the user provides (0 to 10) and NPS score derived (-1 to 1) from the feedback

Parameter	Type	Definition
Score	Integer	A score from 0-10 answering how likely the customer would recommend the product to a friend
NPS Score	Integer	A derived score (between -1 and 1) based on the Score field used in the calculation for NPS

Table 2.2: ***Time Based Features - First Half*** Continuous time intervals between key milestones in customer's experience

Parameter (Abbreviation)	Type	Definition
IKIT Order to Deliver (IOTD)	Integer	The time it takes from when the customer orders the impression kit to when it is delivered to the customer in the mail
IKIT Deliver to Return (IDTR)	Integer	The time it takes from when the user receives the impression kit to when it is sent back to the company lab with the customers teeth impressions molding
IKIT Return to Inspected (IRTI)	Integer	The time it takes from when the user returns the impression kit with the molding to when it is received and accepted into company laboratory to craft a treatment plan against impressions
IKIT Inspected to Treatment Created (IITTC)	Integer	The time it takes from when the impression molding was accepted and a treatment plan is created and visible to the customer for review
IKIT Treatment Created to Aligners Order (ITCTAO)	Integer	The time it takes from when the treatment plan is created and visible to the customer for review to the customer actively making the decision to purchase the aligners based on the treatment plan created

Table 2.3: ***Time Based Features - Second Half*** Second half of continuous time intervals between key milestones in customer's experience

Parameter	Type	Definition
Aligner Order to Dentist Approved (AOTDA)	Integer	The time it takes from when the customer orders the aligners to when the in house dentist reviews the treatment plan and approves it and sends to manufacturing
Dentist Approved to Aligner Shipped (DATAS)	Integer	The time it takes from when the order is approved and sent to manufacturing to when the manufactured aligners and in the mail to the customer
Aligner Shipped to Aligner Delivered (ASTAD)	Integer	The time it takes from when the aligners are shipped to the customer to when it was delivered to the customer
IKIT Order to Aligner Deliver (IOTAD)	Integer	The time it takes from impression kit order to aligners delivered
Length of Treatment (Length)	Integer	The overall length of treatment (proxy based on when the aligners were shipped to the customer to when the satisfaction survey was sent to customer)

Table 2.4: ***Treatment Impacting Features:*** Features that impact a customers experience by either lengthening the treatment itself or delaying the start of the treatment process

Parameter	Type	Definition
IKIT Rekit Needed (REKIT)	Boolean	A boolean value that is true for when a customer's return impression kit is not suitable to craft a treatment and the user must reorder a kit (free of charge) with a clearer mold of their teeth
IKIT Rekit (REKIT_NUMBER)	Integer	For customers where a rekit is needed, the number or times the user had to repeat the ordering process due to an unsuitable molding of their teeth
Entered Stips (STIPS)	Boolean	A boolean value that is true if the user did not complete their online profile (which includes medical/dental history) and is now in a "holding" status until completion of profile. Occurs after user purchases their aligners
Length of Stips (LENGTH_STIPS)	Integer	Should a customer fall in the STIPS status, the overall length the user is in the status (holding) until filling out their online profile
Refinements (REFINE)	Boolean	A boolean value that is true if the user requires adjustment to their treatment plan due to customer discomfort or disatisfaction to how their current treatment plans are going

Table 2.5: ***Cases and Check In Data:*** Features related to the Check In Process and Customer Service Cases

Parameter	Type	Definition
General Cases (G_CASES)	Integer	The number of support cases generated by the customer. Support cases include Lost Aligners, Questions, etc
Pro Care Cases (P_CASES)	Integer	The number cases that ended up requiring support from a special support professional care team to help resolve. These often require higher level of care
Has Clinical Priority Case (CLIN_PRIORTY)	Boolean	A boolean value that is true if a user created a customer support case that needed clinical support from the customer support team. Generally this is for customers in treatment and are expressing discomfort with their aligners
Clinical Priority Cases (C_CASES)	Integer	The number of cases where the type was clinical. This usually is related customers in treatment and are feeling discomfort/ill-fit of their aligners and need attention
Number of Check-ins (Checkins)	Boolean	Throughout course of treatment, the platform asks users to perform a "check in" which asks users a series of questions about their comfort level on their current set of aligners and images of their teeth. The feature appears at a regular interval throughout treatment so customers can perform a check in multiple times over the course. This boolean value that, if true, indicates that the user performed at least one check in throughout their whole course of treatment
Full Checkins (Num_Checkins)	Integer	For those that performed at least check in, this field counts the total number of check ins the customer performed over the course of their treatment.

2.2 Processing

For simplicity on the topic, only records that did not have any missing parameters (detailed under Dictionary) were considered in the dataset. This means that any scores that had a field empty were removed from the dataset. We do this because in the real world setting we would not expect records to be missing as all dates must exist in some data source location in order to proceed in the treatment process. For example, if the time interval it took for a customer to receive an impression kit is empty, then it would not be in included in this dataset. A person must have received the impression kit in order to use the treatment so the missing data is likely to be database related issue. This type of issue will be improved upon assuming the research topic can yield vital information. Therefore, the records removed should be completely at random and should not reveal any bias in our models.

The records were based on survey responses between July 1, 2021 and March 21, 2022. The company that owns the data keeps a repository of scores as well as an identifier related to the account which can tie back to points in their experience. For example, we extracted the continuous time intervals show on table 2.2 and table 2.3 by using the identifier and finding the times of the listed milestones and calculated the time difference. For our categorical features, we also looked through the company's internal database to extract these values using the same account identifier. No personal identifiable information was used in this study.

CHAPTER 3

Exploration of Data

We begin by performing some exploration of the data (EDA) combining both our knowledge of the product flow as well as noting any data points that might correlate to the score. This chapter only outlines some of the data points that saw some indication of relationship to the score and what we ultimately added into our final model.

3.1 Categorical Variable Exploration

One of the first set of features we look into was understanding how our categorical variables impact scores. These features, can be deemed "low hanging fruit" as the fields are interpretable and not to complex (True/False value only). These aspects combined with the fact that these features are actually high impact in terms of a customer's experience seem logical to look into in the beginning. More details will be explained in each subsection.

3.1.1 Refinements

As explained in table 2.4, refinements generally occur when a user must re-perform the impressioning process as the results of the customer's aligner is not matching the treatment plan or there is too much discomfort with the clear aligners beyond the normal range.

Below (Table 3.1), we provide a breakdown of the average NPS scores and the count of customers based on whether or not they were asked to do refinements in their plan.

Table 3.1: ***Refinements Summary:*** Mean score and sample size for customers who underwent Refinements

Refinements	**Mean Score**	**Median Score%**	**Sample Size (n)**
Not Needed	7.81	10	381
Needed	6.86	9	221

Here we see that the average score for customers who complete their treatment without the need for refinements generally score higher on the scale (nearly 14% greater!). We also see the median score higher as well. These basic summary statistics show that Refinements might show a pattern with score.

3.1.2 Stipulations (Stips)

During the initial phases of procuring a treatment plan, the company asks for personal medical and dental information in order to confirm that the customer is physically capable of the clear aligner treatment. Because of this, customers cannot proceed with aligners purchase until after completion of the information intake; a process completed online. This procedure often times delays a customer from beginning treatment as they do not complete the required intake for a multitude of reasons (unaware information was needed, clarity on medical/dental questioned asked/etc).

Table 3.2: ***Stipulations Summary:*** Mean score and Sample size for customers who encountered Stipulations

Stipulations	**Mean Score**	**Median Score**	**Sample Size (n)**
Not Needed	7.50	10	509
Needed	7.29	9	93

Similar to what we saw in refinements, we do see a lower score mean for those that fall into this status but to a lesser extent but the median between the two reflects somewhat

of a shape difference which provides a bit more information about how this feature impacts score. Directionally, this makes sense as customers who now either realized their treatment has been delayed or the fact that they now have an extra step in their process would be frustrated or annoyed thereby providing a lower score.

3.1.3 Clinical Cases

Our analysis on cases data will be given more attention in the latter part of this chapter but to highlight how important understanding cases data is to the score we take a look at the categorical feature of whether a customer opens a Clinical Priority case during their treatment. As a reminder, clinical cases occur when when a customer experiences discomfort or pain above their threshold. These cases generally get routed a special team within the company to try to resolve.

Table 3.3: ***Clinical Priority Cases Summary:*** Mean score and sample size for customers who had at least 1 Clinical Priority Case

Clinical Priority Cases	**Mean Score**	**Median Score%**	**Sample Size (n)**
None Opened	7.83	10	486
Opened	6.19	9	134

The summary results in table 3.3 are not surprising given the severity of the issue. Generally, a customer requesting a clinical priority case generally means that the discomfort/pain they experience is higher than their threshold which undoubtedly would not reflect well with overall experience. Moreover, usually the resolution of a clinical priority case can result in refinements, meaning the patient will need to undergo the impression kit process again for a new treatment plan. Similar to what we've seen, not only is the median lower but also the score mean is about 26% lower than those who did not need to open a clinical priority.

3.1.4 Check Ins

A final categorical feature we look into individually is the concept dubbed "check ins". Throughout the treatment process, the company asks customers to perform a "check in". This essentially allows customers to document their fit and discomfort levels as well as process photos of their progress from different angles of their mouth (top/bottom/side). The flow also allows users to request a customer support representative to reach out in occurrences where fit and discomfort levels are poor.

Table 3.4: ***Check-Ins Summary*** Mean score and sample size for customers who performed at least 1 Check In

Check Ins	Mean Score	Median Score%	Sample Size (n)
Not Performed	7.29	10	382
Performed	7.77	10	220

The check-in feature is one of the only parameters that appears might be positively correlated with the score. While the median appears to be the same in both cases, there appears to be a higher mean score for those that do perform the optional process versus those who don't. This, again, makes sense as this feature might just be a testament to the customer's high level of engagement or their appreciation of such a feature that illustrates the customer's value to the company.

3.2 Treatment Length

We move forward from categorical features into continuous features. Here, we will discuss treatment length from two different aspects; the first is to show how much treatment length impact customer score and the second is to how much impact a large subset of our features indirectly have on score based on the treatment length proxy.

3.2.1 Treatment Length In Relation to Score Response

In our analysis we hope to see some direct relationship to score. In our pursuit to that goal, we look to iteratively find a relationship grouping our scores into 4 levels and observing a boxplot for any correlation.

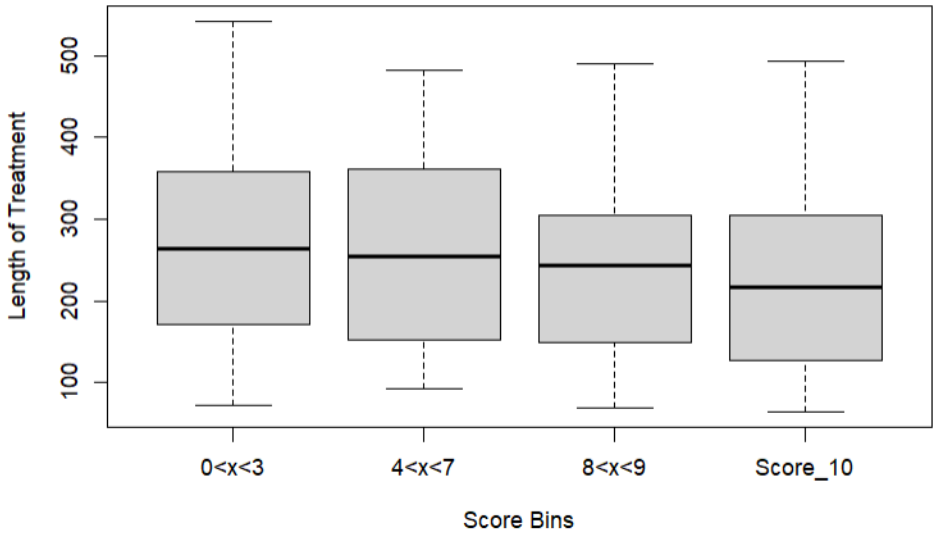

Figure 3.1: Treatment Length vs Score Box-plot Bucketed by Score Bins

Based on our boxplot in figure 3.1, we see signs of a negative correlation when we group scores against length. That is to say that as the length of treatment increases, our score buckets lean towards the lower side. Investigating further with this signal, we fit a logistic regression with the length to the binomial response between a promoter score to detractor/neutral scores. Our goal here is to understand whether length is significant to predicting score and to what degree.

From the regression results we see in table 3.5 [Hla22] which includes the coefficient and standard error, we see that treatment length is significant at $p < 0.01$. Looking at the confidence interval (not shown), we see the results of a good score (9 and up) decreases by .1% to .4% for each additional day of treatment. These results signify that treatment

Table 3.5: Treatment Length Relationship to Score

	Probability of Score>8
Length	-0.003^{***}
	(0.001)
Constant	1.142^{***}
	(0.212)
Note:	*p<0.1; **p<0.05; ***p<0.01

length might be a crucial aspect to our final model. Regardless, however, the indirect impact treatment has with our categorical features also make it necessary to add to our model which we discuss in the next section.

3.2.2 Indirect Relationship to Treatment Length

Something readers will notice in the EDA and final model (chapter 4), is that many continuous variables related to the time intervals between customer touch points is not modelled or shown. A large part of that can be explained that there were no significant relationships we found between score and those time intervals. All is not lost, however. As another step to understanding treatment length, we study how these intervals and other features relates to the treatment length.

Based on the regression results table 3.6, we found a few parameters to be significant to predicting treatment length. One caveat we must mention is the R^2 value. We currently see our R^2 value at .37 (Adj-R^2 at .35). While the value is somewhat low, the key here is understanding the dual benefit of a direct relationship between treatment length and score as well as an indirect relationships with the parameters do not feature in the model proxy

Table 3.6: ***Linear Regression with Treatment Length*** Coefficient and (Standard Error)

	LENGTH
IDTR	2.105*
	(1.256)
REKIT	−40.099*
	(22.061)
ITCTAO	−0.889***
	(0.224)
AOTDA	1.449
	(0.940)
DATAS	2.510***
	(0.581)
ASTAD	7.198***
	(2.441)
STIPS	−23.699**
	(10.232)
REFINE	76.557***
	(9.340)
G_CASES	7.802***
	(1.437)
CLIN_PRIORITY	−39.064***
	(10.263)
Checkins	−39.244***
	(10.390)
Num_Checkins	5.232
	(3.573)
Constant	141.640***
	(13.655)

Note: *p<0.1; **p<0.05; ***p<0.01

through the treatment length.

3.3 Cases

Readers can interpret cases as support efforts opened by the customer. Just as a customer who reaches out to their cable tv provider to ask a question or file a complaint, the clear aligner industry is no different. The difference, however, is the different teams addressing these cases as it ultimately depends on the type of issue (general question or treatment discomfort) and the severity of the issue (pain threshold/etc). Each section below elaborates in detail the type and illustrates the relationship to the score.

3.3.1 General Cases

General cases encompass all known cases opened by a customer. It can range from payment questions, requesting a replacement of clear aligners or even potentially complaints about fit.

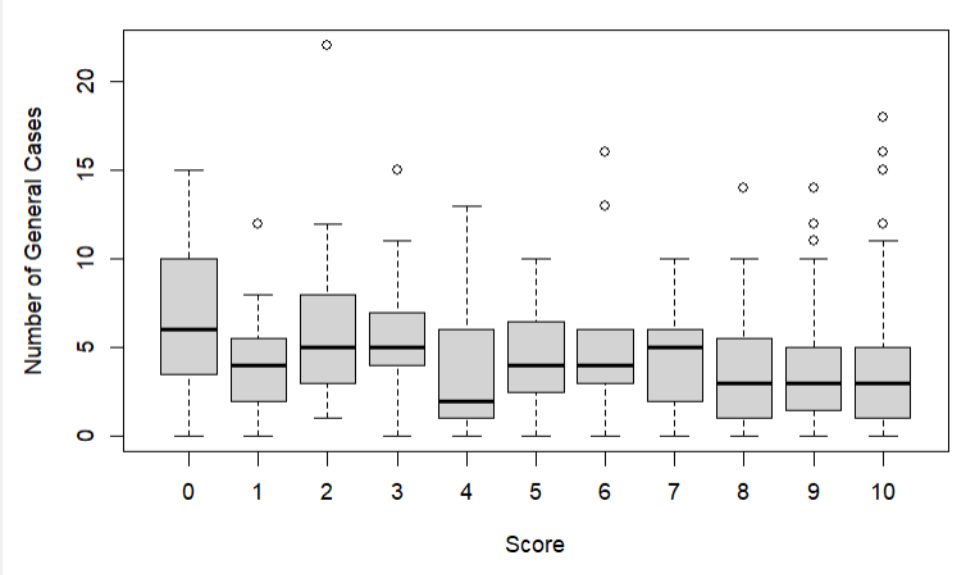

Figure 3.2: Treatment Length vs Score

Using the box plot in figure 3.2, we look into any correlative patterns by score. We notice that as the score gets higher the number of cases start to decline. Looking at the mean cases by score (not shown), we also can somewhat of a linear decline.

3.3.2 Professional Care Cases

Professional care cases are a step up from a general case. At this particular company, there is a division of individuals focused on resolving issues related to this case. Generally, these cases involve the clear aligners themselves and typically related to the discomfort the customer is experiencing. Naturally and luckily, these types of cases are fewer in numbers compared to general cases and a box plot like what was shown in the prior section does not distill too much information. Instead, we plot the mean within each score and fit a line to show the relationships.

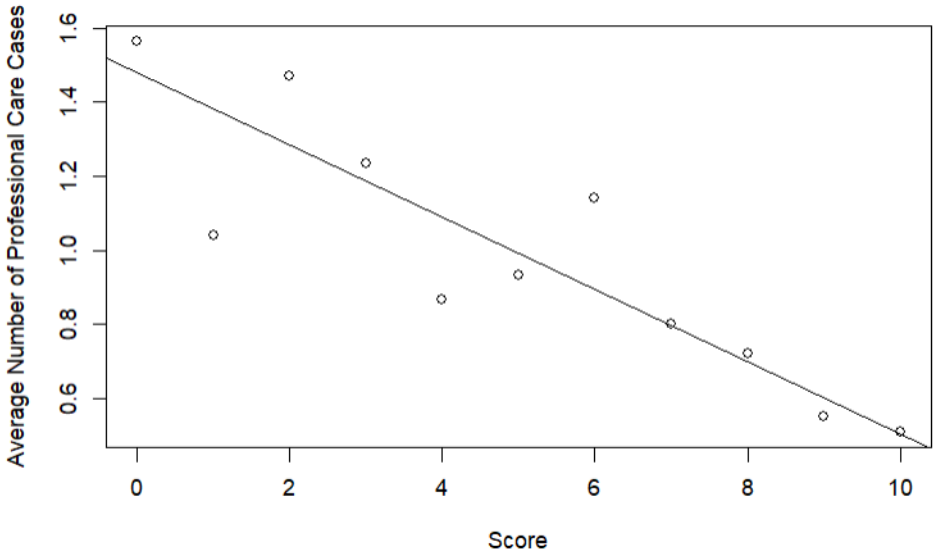

Figure 3.3: Professional Care Cases by Score

As the score increases, the line decreases showing similar pattern as general cases; as the number of professional care cases increases the score decreases. This, too, will be useful in looking at its importance in our model.

3.3.3 Clinical Priority Cases Cases

As referenced under figure 3.4, clinical priority cases are cases that require larger scale attention as the discomfort with the customer's clear aligners have crossed a threshold. Again, luckily the occurrence of this is even less so than professional care cases.

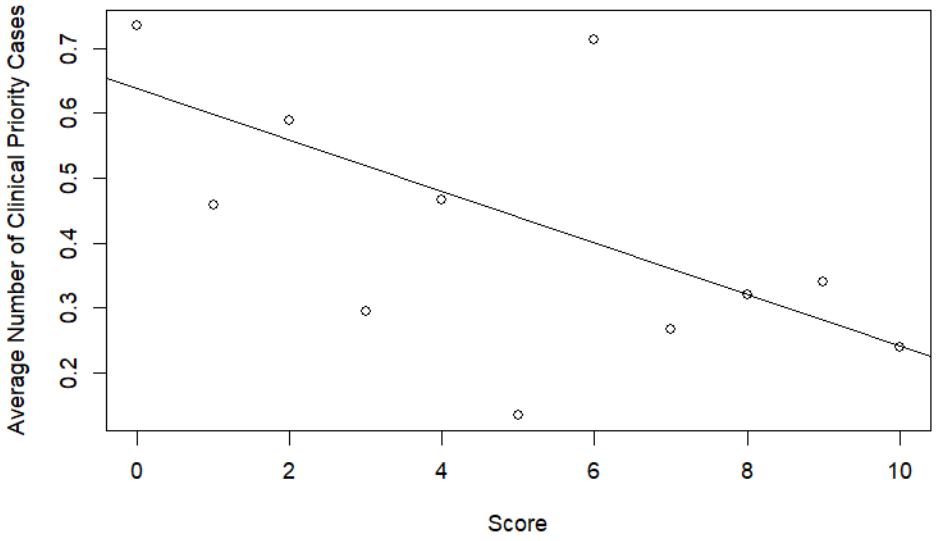

Figure 3.4: Clinical Priority Cases by Score

Again, we see the same correlations we have been seeing with professional care cases and clinical priority cases with increasing cases and lower scores. We will ultimately try adding all cases related information in our model and observe results.

3.4 Interaction Variables

In our data set we have at least 2 categorical features which gives us the opportunity to understand interactions variables. Below we test a series of interaction variables.

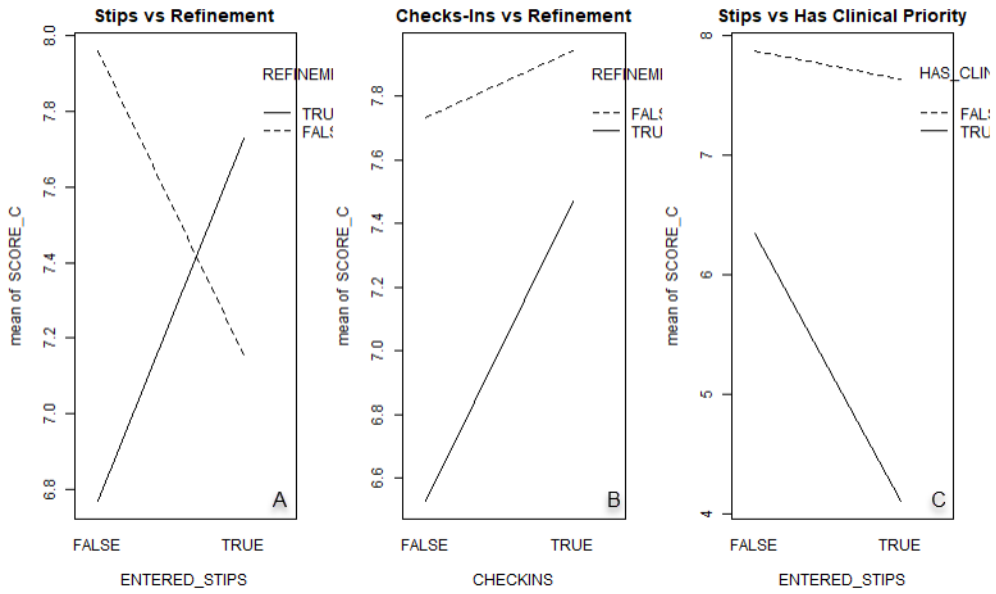

Figure 3.5: ***Iteration Plot -*** Plot A represents interaction between STIPS and REFINE and it appears to have the strongest interaction relationship. Plot B and C represent interaction between Checkins and Refine and STIPS and CLIN_PRIORITY respectively

3.4.1 Plots

On figure 3.5, we notice a series of interaction variables that might be helpful to the model. There were a series of other interactions that also had signs of interaction but none that were stronger than the plots we see nor did any of them make sense in having interactions between each other. To provide a bit more context, interaction effects is when one variable is dependent on the value of another [Fro22]. Generally, when one variable affects the outcome of the dependent variable, we would call that a main effect. In a more complex regression, we can implement interaction variables to understand how might an additional variable impact the original independent variable and dependent variable. To help illustrate if two variables have an interaction effect, we use an interaction plot by which when one line appears to cross over with the other line (like we see in figure 3.5) then further investigation is necessary on whether their interactions have a significant effect on the model. One that appears to be

dominant is the cross relationship between Stipulations and Refinement out of the three plots we illustrate. This would mean that the NPS Score have an interaction between Stips and Refinement where the result would depend on the combination of the two variables.

3.5 Transitioning to a Model

It appears that we have a mixture of both categorical variables and continuous variables to try to find a model that has a strong predictive power and is interpretable. To recap, we will try to use the following:

1. Categorical Data (True/False) from Refinements, Stipulations, Clinical Priority Cases Opened, Check Ins Completed
2. Treatment Length
3. Cases Data Severity and the number of cases opened
4. Interaction Variables

In the next chapter, we will discuss the models we ultimately denote as our final models based on the features we found might showing strong relationships with Score.

CHAPTER 4

Statistical Modeling

Based on our data exploration, we attempt to fit those features that seem to show signs of a relationship to score in order to satisfy our research objective. We end up with two final models; the first being logistic regression while the second being ordinal logistic regression. Notice that we employ two types of regression and not machine learning models (Tree Based Classifiers and the like). More discussion about those usages will be covered in latter chapters but the overall intent of our research is to understand which features impact score and to what degree; this is an aspect machine learning models generally cannot address.

This chapter will cover the results of those models in accuracy with the interpretation discussed in the chapter following.

4.1 Logistic Regression

4.1.1 Intro to Logistic Regression

Logistic Regression is a statistical model that can be used to predict the probability of a binomial outcome. In regression analysis, the use of logistic regression uses the parameters (features) of our dataset to estimate its value in our dependent variable. In short, the regression predicts the outcome by taking the log-odds (logarithm of odds) for the event based on the linear combinations of one or more independent variables. The logistic regression is a sigmoid function that takes in an input (t) and outputs a value between 0 and 1. The function takes on the model form below:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-\beta^T x}}$$

where β represents a matrix of coefficients for the parameters and P(x) can be defined as the probability of an event occurring given some (x). Then, to more easily understand the interpretation of the function, we take the logit (log odds) functions which is the inverse of σ and end up with the following:

$$\frac{p(x)}{1 - p(x)} = \beta^T x$$

To put it simply, we are able to understand how each independently impacts the probability of certain binary outcome occurring [Fou22b] [Far06]. Figure 4.1 illustrates a basic example of a logistic regression.

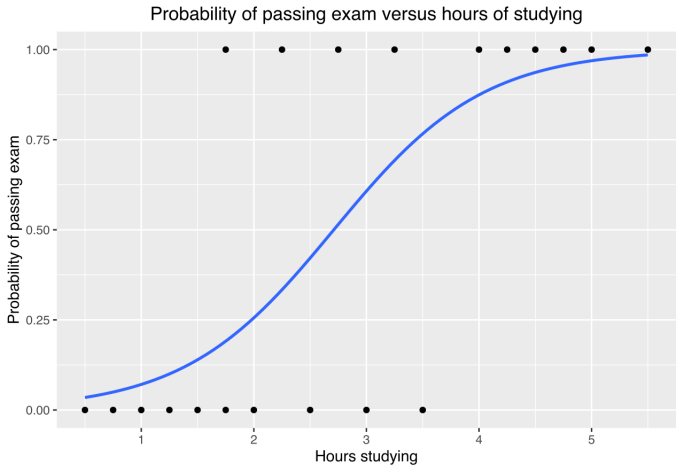

Figure 4.1: Simple Logistic Regression Example [n22]

4.1.2 Logistic Regression Applied

If readers recall, our overall research goal is being able to understand and predict customer response scores on their product experience. The purpose of the logistic regression is to be able to understand factors that are important to a binary outcome. Thus, we want employ

the logistic regression to understand how well a logistic regression can predict a detractor (poor) score from a neutral/promoter (good) score. Again, recall that a detractor in NPS is when a customer responds with a score equal to or below 6. Meanwhile a neutral/promoter score is when a customer responses with a score higher than or equal to 7. This is where we perform our cutoff. Anything below a score of 6, we denote it's dependent variable to be 0 and any score above is 1. Using what we learned in chapter 3, we begin by fitting a model for the logistic regression with the features we found under Section 3.5.

4.1.3 Results and Accuracy

We will discuss in detail the interpretation and significance of the features we use in the next chapter, but here we employ the features we found to show relationship towards score and the output of the results are show in table 4.1. Below under table 4.2, we used the leave one out cross validation (LOOCV) to understand the model accuracy against the dataset.

Table 4.2: Confusion Matrix: Logistic Regression Accuracy

	Reference Negative	**Reference Positive**
Prediction Negative	55	60
Prediction Positive	111	376

Overall, we obtain an accuracy of about 72%. This means that our models with the features we included was able to correctly predict about 431 out of 602 records. Some would argue this would be a solid prediction accuracy. One caveat, however, is our true detractor score (sensitivity). Our model was able to predict a negative score about 33% of the time and it was able to predict and true neutral/positive score about 86% of the time. The good news here is that there is a model that has some predictive performance but we hope with the ordinal regression and our removal of outliers can further refine our model.

Table 4.1: ***Logistic Regression - 6 Score Cutoff*** Coefficient and (Standard Error)

	Probability of Score > 6
STIPS	−0.659**
	(0.302)
REFINE	0.436
	(0.290)
LENGTH	−0.002
	(0.001)
G_CASES	−0.084**
	(0.040)
P_CASES	−0.539***
	(0.151)
CLIN_PRIORITY	−0.520
	(0.398)
C_CASES	0.569**
	(0.247)
CHECKINS	0.333
	(0.214)
STIPSTRUE:REFINE	0.978
	(0.645)
Constant	1.892***
	(0.296)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.2 Ordinal Regression

4.2.1 Intro to Ordinal Regression

We move forward from the logistic regression to use a regression that seems to make more sense to apply to our research case; ordinal logistic regression. Ordinal Logistic Regression [Par16] is a statistical analysis method to understand the relationship between an ordered response variable and one or more explanatory variables. The key here is that the intention of the regression is it maps in responses where there is a clear order. Additionally, the regression can also take in categorical and continuous variables. The ordinal logistic regression is essentially an extension of the logistic regression (and interpretation is similar) but instead of mapping to a binary response (2-levels), the regression can map to multi-level responses (k-levels).

Focusing briefly on the theory behind it [n], let's have a dependent variable (Y) be the outcome with (t) categories (i.e. 0 to 10 or -1 to 1). The probability of some (Y) being less than or equal to a category defined in t is defined as.

$$\frac{P(Y \leq t)}{P(Y > t)}$$

Similar to the logistic regression, we can also get the log odds form (logit) by completing the following:

$$log\frac{P(Y \leq t)}{P(Y > t)} = logit(P(Y \leq t))$$

Given the similarities with logistic regression, we focus more closely on how it's applied to our research.

4.2.2 Ordinal Regression Applied

Recalling from chapter 2, companies use a feedback score ranging from 0 to 10 that ask the likelihood a customer would recommend the product to a friend/colleague. We could have used that as our dependent variable but a couple things may make this difficult. The first is that fitting a model from 0-10 involves 11 categories which is somewhat complex given that we've only just introduced this research topic. Secondly, predicting a score between a 0-10 scale doesn't carry much meaning in telling how well the product does. For example, we understand a customer rating a 0 means the product was absolutely terrible and that a score of 10 means the product was great but what does it mean when a product scores close to 3 or how much more valuable is it if they score a 4 instead? We decide to forgo using the 0-10 score as our response and use the NPS based scoring defined under table 2.1 as it more simply gauges consumer sentiment.

Like we covered under Section 1.3, the final net promoter score can be bundled into 3 distinct categories; promoter, neutral and detractor. Those 3 categories are predicated upon the 0-10 scale score and it simply categorizes customer sentiment either positive/unbiased/negative. Applying the ordinal regression keeps track of the fact that the responses are ordinal but still allows us to understand what specific features in our model are contributing to the score. Some may argue that there could be some sense in using the ordinal regression to predict on the 0-10 scale but one of the key figures in the final NPS score is actually the formula provided below:

$$FinalNPSScore = \frac{Promoters - Detractors}{TotalRespondents}$$

Thus, predicting the 3 level response is less labor intensive and arguably more beneficial to our research topic. Figure 4.2, reviews the different levels of NPS scores and what they mean.

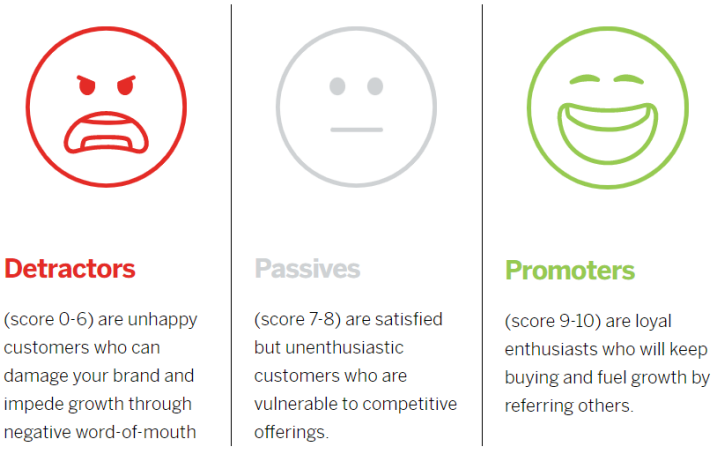

Figure 4.2: General Interpretation of NPS Scores [Qua22]

4.2.3 Results and Accuracy

Again, we discuss the implications and significance of the features we map in the next chapter but we employ the same features as what we did with the logistic regression. One drawback of the ordinal logistic regression is that results from the regression aren't as easily interpretable compared to that of the logistic regression. The table of coefficients 4.3 and the accuracy using LOOCV under Table 4.4 are outlined below:

Table 4.4: Confusion Matrix: Oridinal Logistic Regression Accuracy

	Reference Detractor	**Reference Neutral**	**Reference Promoter**
Prediction Detractor	32	10	34
Prediction Neutral	0	0	0
Prediction Promoter	134	52	340

In reviewing the accuracy, overall we achieve an accuracy of 62% with our model having a strong predictive power for true promoter scores (91%) but lacking in the other two levels with a 19% accuracy in true detractors and no guesses for neutral scores. As a reminder, however, is that we are predicting 3 levels and it is expected that accuracy will go down compared to logistic regression. The fact that we were able to somewhat able to accuracy

Table 4.3: Ordinal Logistic Regression - NPS Score

	NPS_SCORE
STIPS	-0.565^{**}
	(0.269)
REFINE	0.451^{*}
	(0.267)
LENGTH	-0.001
	(0.001)
G_CASES	-0.074^{*}
	(0.038)
P_CASES	-0.588^{***}
	(0.144)
$\text{CLIN}_{P}RIORITY$	-0.104
	(0.385)
C_CASES	0.481^{**}
	(0.238)
CHECKINS	0.333^{*}
	(0.189)
STIPSTRUE:REFINE	0.648
	(0.545)
Note:	*p<0.1; **p<0.05; ***p<0.01

predict one level of NPS score well shows promise and in the next section we cover how we might refine the model through clustering.

4.3 Clustering

In the past few chapters and sections, we have done a large amount of heavy lifting; we explored what variables might have a solid relationship with score, modeled the data to a couple regression models and worked through their respective accuracy. Despite these steps, we are not quite satisfied with our level of accuracy although some may argue that having a low predictive power is better than no model at all. Nevertheless, as a final attempt to improve our accuracy we try to apply a clustering method to see if we can group our records with similar features and dissect records that do not fit our scope of research due to a feature strongly correlated with the response resulting in an outlier.

4.3.1 Applying K-Means Clustering

We will make use of the K-means Clustering [Gar18] model against our records to see if we can isolate and remove outliers. We begin by looking at our confusion matrix from Table 4.4. Notice that we have an overabundance of incorrect predictions for both true detractors and true promoters so we look into clustering both of those subset separately.

Before clustering, we must get insight as to how many clusters we want to form to optimize and distill the most amount of information. To achieve that, we utilize the elbow method and plot the within sum of squares against the number of clusters and find "the elbow" at which the bend occurs. That elbow would indicate the best number of clusters to use in K-means. Figure 4.3, illustrates the optimal cluster numbers in our true promoters which is 3. Our detractor coincidentally (not shown) also has 3 cluster thus we will be dissecting our data based on 3 centroids for both.

In choosing what features we want to look into, it would likely be best to look into

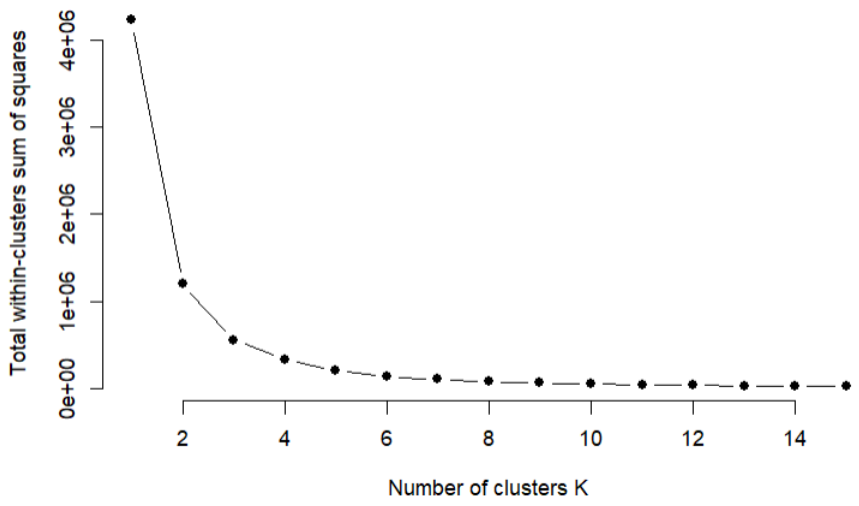

Figure 4.3: Optimal Number of Clusters (True Promoters)

continuous variables. Given our categorical are primarily composed to two levels, it would be hard to distill any information due to the lack of variation in the feature. Figure 4.4 illustrates the box plot based on the 3 clusters we created against the treatment length.

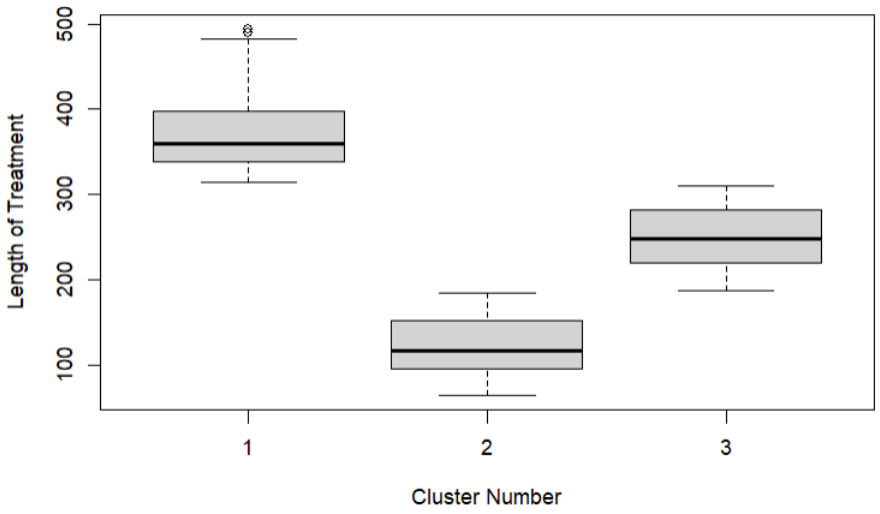

Figure 4.4: Boxplot: Treatment Length vs Cluster (Promoter)

Recall from both our ordinal and logistic models from Chapter 4, that there seems to be an inverse affect on treatment length to NPS score. That is to say, for every extra day a

customer remains in treatment, the probability of a positive score decreases. Looking at the boxplot of promoters scores, we notice cluster 1 has a high length of treatment compared to cluster 2 and 3. This begs the question of why, despite a high treatment length, would customer's score result in a promoter. We conclude, that this must be an outlier and remove the records (88 total) from our dataset.

We now transition to our true detractor score and apply the same methodology with treatment length. Looking at figure 4.5, we notice the same phenomenon except in the other direction; cluster 1 despite a negative correlation still gives a detractor score despite a low treatment length. In light of this, we, again, denote this as outliers and remove the records (52 total).

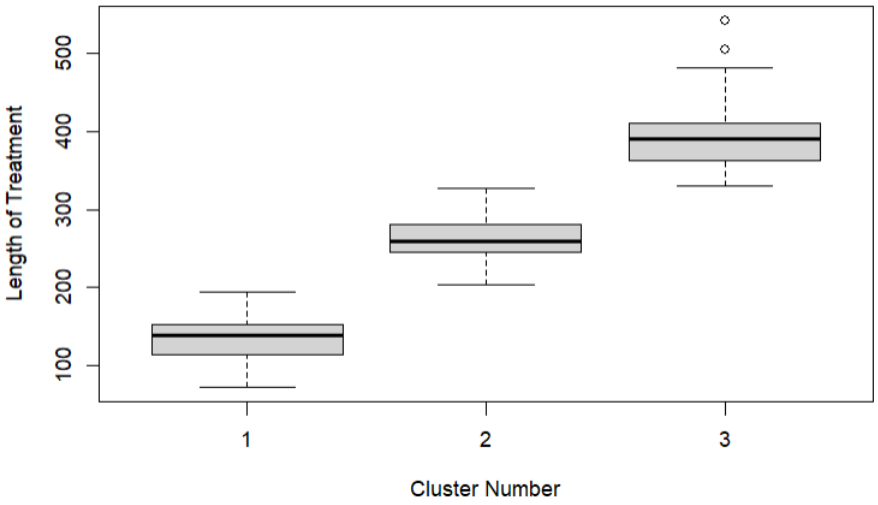

Figure 4.5: Box plot: Treatment Length vs Cluster (Detractor)

Based on this removal of outliers, we still have ample sample size (462) and we can proceed with rerunning our models in hopes of more accuracy and information of our features.

4.3.2 Results

Starting from the logistic regression and table 4.5, we already see some improvement with the model itself with more features in the model moving into significant territory. Again, we will cover the interpretation in the next chapter and we focus on the accuracy. As a reminder, our logistic regression is meant to understand whether we can predict a poor score from a neutral score. Our confusion matrix shows an overall accuracy of 84% (a 12% points increase!). This increase is lead by improvements in predicted detractor and predicted neutral/promoter scores as the accuracy is 68% (compared with 33%) and 90% (compared with 86%). Overall, we can argue that it is wise to exclude those outliers and we will discuss more on possible alternatives we can do instead of simply removing the records.

Table 4.6: Confusion Matrix: Logistic Regression Accuracy

	Reference Negative	**Reference Positive**
Prediction Negative	78	36
Prediction Positive	36	312

Applying the same to our ordinal logistic regression model, we, again, do see more features falling into significant territory. We look at the confusion matrix and also see an improvement in our overall accuracy of 74% (up 13% from our original dataset). These increases are also driven by an increase in accuracy of our predicted detractor scores of 63% (up 44% points) and true promoters score of 94% (up 3% points). Unfortunately, however, our model still is not able to predict any neutral/passive scores. Despite this, we will discuss why this might be happening and why we would still use this model in the next chapter.

We have gone nearly end to end in the intensive exploration analysis and modeling process. In the next chapter, we will discuss the importance of the features we use in the model as well as other models we considered.

Table 4.5: Logistic Regression - NPS Score (Outliers Removed)

	Probability Score > 6
STIPS	−0.046
	(0.565)
REFINE	0.281
	(0.375)
LENGTH	−0.016***
	(0.002)
G_CASES	−0.188***
	(0.063)
P_CASES	−0.439**
	(0.214)
CLIN_PRIORITY	−0.644
	(0.537)
C_CASES	0.687**
	(0.344)
CHECKINS	0.417
	(0.323)
STIPSTRUE:REFINE	0.898
	(0.983)
Constant	6.305***
	(0.659)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4.7: Ordinal Logistic Regression - NPS Score (Outliers Removed)

	NPS_SCORE
STIPS	−0.178
	(0.378)
REFINE	0.406
	(0.318)
LENGTH	−0.013***
	(0.001)
G_CASES	−0.144***
	(0.051)
P_CASES	−0.594***
	(0.191)
CLIN_PRIORITY	0.105
	(0.478)
C_CASES	0.608*
	(0.320)
CHECKINS	0.305
	(0.245)
STIPSTRUE:REFINE	0.574
	(0.688)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4.8: Confusion Matrix: Ordinal Logistic Regression Accuracy

	Reference Detractor	**Reference Neutral**	**Reference Promoter**
Prediction Detractor	72	19	17
Prediction Neutral	0	0	0
Prediction Promoter	42	43	269

CHAPTER 5

Discussion

As we discussed in the last chapter, we created a model that is accurate and arguably could be used in real world setting without major modifications. That, in and of itself, is an incredible feat and we highlight the reasons in our conclusion. But one of research goals was to understand which features impact the score and to what degree which is why we preferred fitting our models through logistic regression. In the sections below we breakdown both the logistic and ordinal logistic regression in their coefficients, confidence bands and a commentary from a business perspective on why the results we see makes sense. Of course, we did not simply focus on the two models throughout the entirety of this research. We also looked into other machine learning models and provided the accuracy as well as an explanation on why the logistic/ordinal regression was preferred. As a note to the readers, for simplicity and to reduce repetitiveness we compare our models against the removed outliers case (Section 4.3) and the three leveled response (promoter/neutral/detractor).

5.1 Logistic Regression

5.1.1 Interpretation of Coefficients

Referring to table 4.6, we have found the length of treatment, general number of cases, professional care cases, clinical priority cases to be significant in our set of features. We, however, leave in the insignificant parameters as it provides a bit more information on the pattern that we see. Table 5.1 below looks at the odd ratio (for simplicity) and the

Lower/Upper Bound Confidence Bands with a $\alpha = 0.05$:

Table 5.1: Logistic Regression: Odds Ratio and Confidence Interval

Feature	**Odds Ratio**	**2.5% LB Conf**	**97.5% UB Conf**	**Interpretation of Coefficients**
Enter Stips (True)	0.9549768	0.3301874	3.1053353	If Ccstomer Enters Stipulation, the odds of an NPS Score>5 decreases by a factor of .05
Refinements True	1.3247464	0.6409625	2.8026822	If customer must complete refinements, the odds of an NPS Score>5 increases by a factor of 1.3
Length of Treatment	0.9836644	0.9793787	0.9875346	Each additional day in treatment lowers the odds of an NPS Score>5 by a factor of .017
Number of General Cases	0.8287570	0.7306155	0.9355065	Each additional general case opened lowers the odds of an NPS Score>5 by a factor of 0.17
Number of Professional Care Cases	0.6446922	0.4181937	0.9722896	Each additional professional care case opened lowers the odds of an NPS Score>5 by a factor of .0.36
Has at least 1 Clinical Priority Case	0.5251596	0.1827061	1.5113540	Having at least one clinical priority case decreases the odds of an NPS Score>5 by a factor of .48
Number of Clinical Priority Cases	1.9869721	1.0275485	3.9977460	Each additional clinical priority case opened increase the odds of an NPS Score>5 by a factor of 1.9
Check Ins True	1.5181093	0.8117665	2.8958793	Performing at least one check in increases the odds of an NPS Score>5 by a factor of 1.5
Stips (true) x Refinements (true)	2.4549251	0.3730915	18.4574068	A customer with both Stipulations and Refinements increases the odds of an NPS Score to the next level (0 or 1) by 2.4

Starting with length of treatment, for every 1 unit increase there is a .02 chance decrease of a neutral/promoter score. Directionally, this makes sense as customers that face long periods of treatment length would likely need to endure discomfort or, at the very least inconvenience, which slowly wears away at customer's patience.

Moving on to cases, in both number of general and professional care cases created, we see a decline of .17 and .36, respectively. Again, cases are generally opened when there is a general question or when there is discomfort with the aligners so we'd expect to see a lower score driven by the inconvenience as well as the root cause of why customers would be opening a case. Additionally, we see the severity of the decline line up as well in that general cases decline of a better score is less in weight compared to the more severe professional care cases.

Finally, in our last significant variable, our odds of a higher rate score increases for each clinical priority case opened. This feels counter intuitive to what we'd expect based on what we see in the earlier two case categories. Clinical Priority Cases is likely the most severe of cases so instinctively we should see the odds in the opposite direction and similar to professional care and general cases. More research should be reviewed into this. Some explanation of this unexpected trend include the level of care the clinical priority team members take in supporting these customers. Notice, however, our categorical feature of having at least 1 clinical priority case our observed odds ratio shows a large decline in odds of a better score rating. This furthers the notion that it is perhaps more research needed to understand the true direction of clinical priority cases and it's true implication to score.

A couple honorable, but insignificant, mentions in our set of features is our "checkins true" parameter and our "enters stips". Here, we see the odds ratio increase by .52 when Check Ins are true. This appears to make sense as Checks Ins allows customers to stay engaged with the treatment and gives customers assurance that the company truly cares of the outcome of results. Stips, on the other hand, is somewhat of an inconvenience to the customer as it potentially leads to delays in starting treatment due to the lack of information

provided by an unengaged customer. Figure 5.1 below visually summarizes our interpretation above via an odds ratio plot (forest plot).

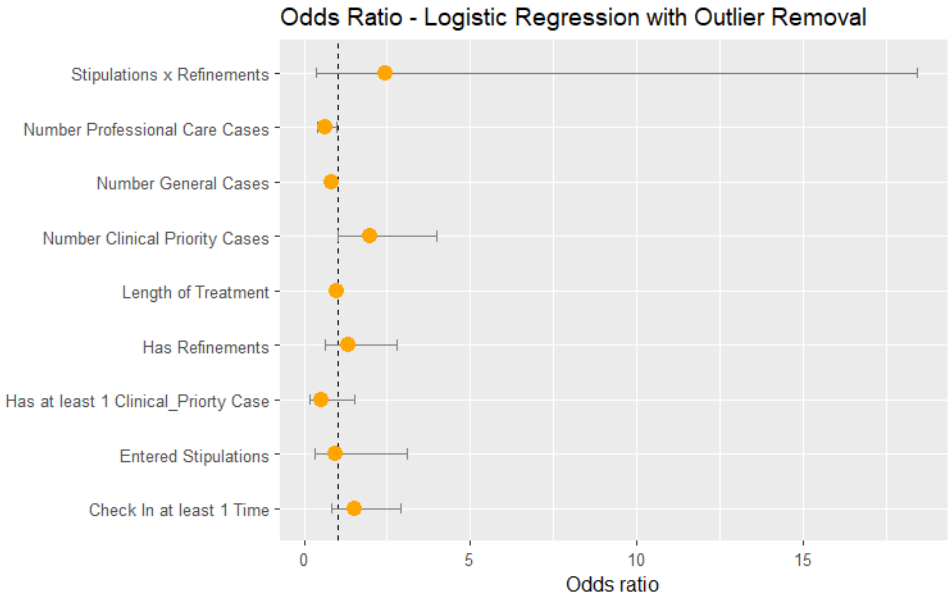

Figure 5.1: Odds Ratio Plot (Forest Plot) - Logistic Regression

5.1.2 Commentary of Results

Generally, our coefficients (even the insignificant features) make sense from the perspective of the business. With the exception being "Number of Clinical Priority Cases", we would anticipate the direction the coefficients play in the likelihood of a positive/negative score. Additionally, when we compare the deviance, our null deviance comes out to be 516.28 and our residual deviance results in 311.69 which shows our model has a better fit and is adequate.

5.2 Ordinal Logistic Regression

5.2.1 Interpretation of Coefficients

As mentioned in Section 4.2, the ordinal logistic regression is an extension of the logistic regression and because of that it would make sense that we will see results, in the form of odds ratio and confidence intervals, to be similar.

Table 5.2: Ordinal Logistic Regression: Odds Ratio/Confidence Interval/Interpretation

Feature	**Odds Ratio**	**2.5% LB Conf**	**97.5% UB Conf**	**Interpretation of Coefficients**
Enter Stips (True)	0.8370250	0.4050608	1.7967268	If customer enters Stipulation, the odds of an NPS Score to the next level (0 or 1) decreases by a factor of .16
Refinements True	1.5001901	0.8102232	2.8294798	If customer must complete refinements, the odds of an NPS Score to the next level (0 or 1) increases by a factor of 1.5
Length of Treatment	0.9873371	0.9843721	0.9901436	Each additional day in treatment lowers the odds of an NPS Score to the next level (0 or 1) by a factor of .013
Number of General Cases	0.8662310	0.7828610	0.9567449	Each additional general case opened lowers the odds of an NPS Score to the next level (0 or 1) by a factor of .0.14
Number of Professional Care Cases	0.5521332	0.3746716	0.7942238	Each additional professional care case opened lowers the odds of an NPS Score to the next level (0 or 1) by a factor of .0.45
Has at least 1 Clinical Priority Case	1.1106151	0.4386440	2.8712202	Having at least one clinical priority case increases the odds of an NPS Score to the next level (0 or 1) by a factor of 1.1
Number of Clinical Priority Cases	1.8365256	0.9785610	3.4581547	Each additional clinical priority case opened increase the odds of an NPS Score to the next level (0 or 1) by a factor of 1.8
Check Ins True	1.3566859	0.8423970	2.2009601	Performing at least one check in increases the odds of an NPS Score to the next level (0 or 1) by a factor of 1.4
Stips (true) x Refinements (true)	1.7751631	0.4662383	7.0563753	A customer with both Stipulations and Refinements increases the odds of an NPS Score to the next level (0 or 1) by 1.7

Referring to table 5.2, we do see that results trend similar to the logistic regression; our parameters that we saw were significant are still significant in this model and the coefficients signs line up when deducing business logic. The table also includes an interpretation the coefficient (note that not all parameters are significant). Something to note is that our ordinal regression is a little bit harder to interpret compared to the logistic. A large part of that is due to logistic showing a binary response where as ordinal can have as many as k-levels. Again, this is why we only have the 3-level NPS response score instead of 0-10; for simplicity purposes. Nevertheless, the interpretation is somewhat similar but this time we are compare between two of the three models at a time (i.e. more/less likely to give a promoter score than a neutral score or neutral score over detractor score). As such, our commentary and interpretation of the coefficients should not really change from a conceptual point of view with the logistic regression.

Our treatment length feature shows a .13 unit decline for each unit increase in days of treatment. This is down just slightly in comparison to our logistic regression. Moving on to general cases and professional care cases, we see for every additional cases opened, we noted a .13 unit decline and a .45 unit decline, respectively. Finally, our coefficient in clinical priority cases (one of the most severe of cases) increases our changes of a better score by a factor of 1.99 holding all over variables constant. Again, this is likely something that needs to be investigated before we conclusively declare this to be true. To avoid repetitiveness, our sentiments about our other honorable yet insignificant features still hold and we still feel that it holds value to leave in the model especially when talking about further research into the model that we will discuss as a future improvement item in the next chapter. Additionally, the odds plot ratio of figure 5.2 illustrates the significant and not significant features and it's impact on the model.

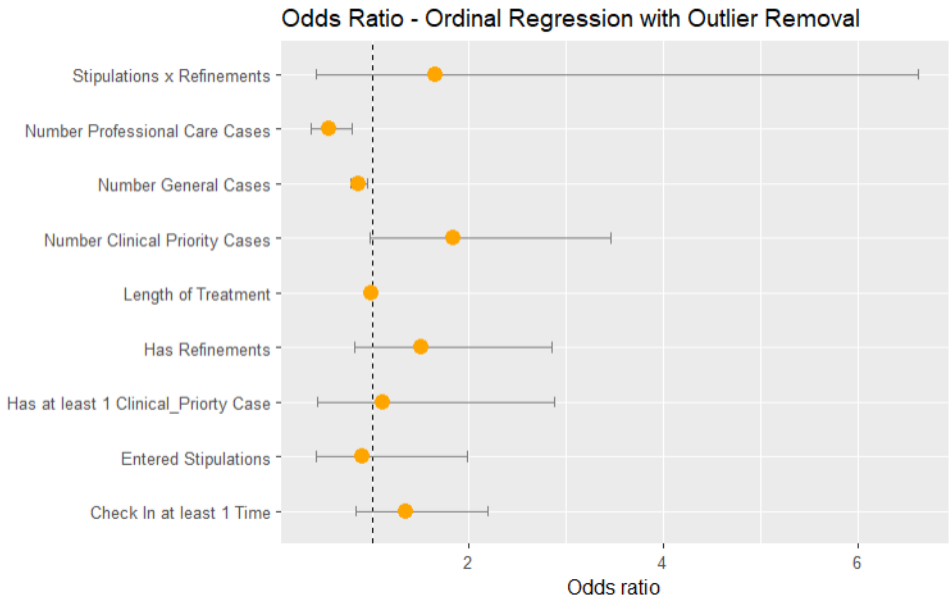

Figure 5.2: Odds Ratio Plot (Forest plot) - Ordinal Regression

5.2.2 Commentary of Results

The results do not change much comparing between a logistic from ordinal regression. Some coefficients become more muted in terms of odds ratio while others become a bit more impactful. Our research goal is meant to set the foundation of a predictive model placed in real world settings. With that in mind, the ordinal regression model seems to be the wise choice in proceeding towards that goal. Despite that model unable to predict a neutral score, we still achieve an agreeable accuracy rate and leave the door open for model refinements which we discuss some of the ideas in our conclusion.

5.3 Other Types of Modeling Methods

The following sections will give some insight why the ordinal and logistic regressions were the final models we selected. Many of the models we mentioned are classifier based and therefore would be a bit more difficult to interpret and even more difficult to refine and improve upon.

As a reminder, we focus primarily on the outliers removed scenario and the 3-level response.

5.4 Naive Bayes

5.4.1 Intro

Naive Bayes is a probabilistic classifier machine [Gan18] learning model that is specifically geared to classify the data set. One of the main aspects of Naive Bayes hinges upon the Bayes Theorem which essentially states what we can find out probability of an event happening based on prior knowledge in relation to the event (refer to figure 5.3 below). One of the assumptions about the Naive Bayes classifier is that the predictors/features are independent. Based on earlier exploratory analysis we did meet that criteria.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 5.3: Bayes Theorem [Fou22a]

5.4.2 Results and Commentary

Based on table 5.4, overall we have a respectable 71% accuracy. Within each level we see the model was able to predict 63% of our true detractors, 3% of true passive (neutral) and 88% of our true promoters. This actually comes relatively close to our ordinal logistic regression model which ended up close to 74% accuracy overall. Moreover, the classifier model was actually about to predict a few neutral scores.

Table 5.3: Confusion Matrix: Naive Bayes Accuracy

	Reference Detractor	**Reference Neutral**	**Reference Promoter**
Prediction Detractor	72	16	31
Prediction Neutral	7	2	1
Prediction Promoter	35	44	254

5.5 Random Forest

5.5.1 Intro

Our next classification model, random forest, focuses on creating a large number of decision trees during the training process. The crux of the random forest model is that the classification of the output is decided upon the most "votes" based on the number of trees that were created in an uncorrelated fashion [Fou22c][Dav20] (see figure 5.4).

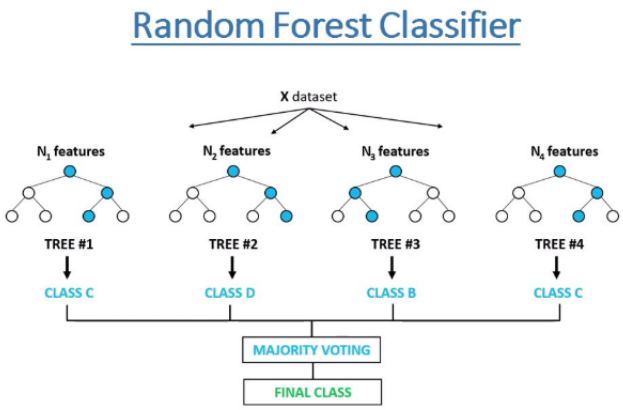

Figure 5.4: Random Forest [Cha21]

5.5.2 Results and Commentary

The Random Forest overall accuracy ended up at 66% with true detractors at 39%, true neutral at 46% and 80%. The classifier model was actually able to predict more neutral

values but it seems to have come at a cost at the accuracy of the other two levels. In fact, it seems like the model predicted the neutral value nearly a third of the time. This model might be insufficient as it almost seems as if it's predictions are arbitrary.

Table 5.4: Confusion Matrix: Random Forest Accuracy

	Reference Detractor	**Reference Neutral**	**Reference Promoter**
Prediction Detractor	45	6	2
Prediction Neutral	58	29	55
Prediction Promoter	11	27	229

5.6 Support Vector Matrix

5.6.1 Intro

The last model we look into is the Support Vector Matrix (SVM). SVM, to put it simply, is a classifier that takes data points and creates a hyperplane known as the decision boundary [Le18]. Depending on the complexity of the data point, the decision boundary can be linear, circular, etc. The goal of the SVM is to create decision boundary that divides up the points such that it minimizes the margin of error between the boundary to the points.

5.6.2 Results and Commentary

Our SVM results was similar to our ordinal regression in that it was not able to make neutral prediction. Overall our accuracy of the model was at 75% with a true detractor accuracy of 69% and a true positive accuracy of 93%. One thing to note was the length of time it took the model to run. Using SVM is a double edged sword in that the SVM model does have more levels of adjustments such as kernel type and the tuning parameter. Depending on the selection of these can ultimately impact the accuracy which can shorten or lengthen the computation time. The time it took for this model, however, was quite long which may

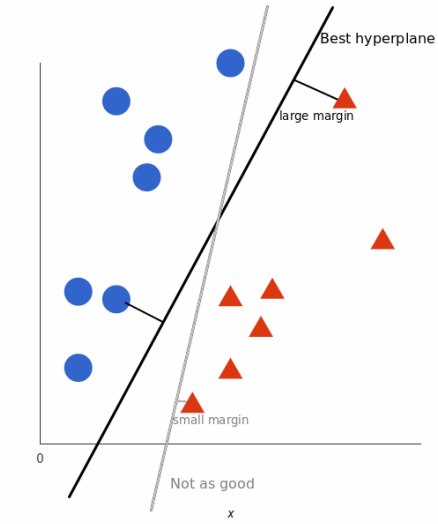

Figure 5.5: Support Vector Machine [Le18]

be a problem if implementing at scale was a deciding factor.

Table 5.5: Confusion Matrix: Support Vector Matrix

	Reference Detractor	**Reference Neutral**	**Reference Promoter**
Prediction Detractor	79	19	19
Prediction Neutral	0	0	0
Prediction Promoter	35	43	267

5.7 Preferred Model Outcome

One general underlying theme across all the classifier models we introduced in this chapter is that all were unable to effectively quantify how much each feature in the model are valued. For example, in both the logistic and ordinal regression we were able to use the odds ratio to determine both the direction the features played in predicting the overall score and to what degree and significance. The only model that can do something close to this would be the SVM model but it only illustrates the importance of each feature in the overall model.

Going into each model individually, we rule out the SVM model as it was computationally

expensive. In real world applications, particularly with a larger subset of data, we need to strike a balance between accuracy and performance. In this instance, we were able to achieve close to the same amount of accuracy as the ordinal regression model but achieving this level of accuracy was quite time intensive.

Moving onto the Random Forest Classifier, admittedly one of the drawbacks of either the final models is that it does not have (yet) the ability to predict neutral results which the random forest model does. The random forest model, however, seems to be taking somewhat arbitrary guesses as nearly a third of it's predictions are for neutral and the number of true neutral results are far less than it's prediction. Given we have the ability to understand the coefficient in ordinal regression it makes it a bit easier to refine the model over random forest.

Finally, naive bayes appears to be the closest competitors to our full model. It matched equally accurate against our true detractors and even had some predictions of neutral classifications. In the end, picking the ordinal regression was the right choice as it could predict true promoter marginally better and referring back to our original point in being able to quantify our features.

CHAPTER 6

Conclusion

6.1 Research Conclusions

We now move into our conclusion which discusses things we learn, next steps for improvement and illustration of importance of this research topic.

6.1.1 Lack of Impact from Early Stages Experience

One of the major revelations we have found in this research is that early customer experience has no direct impact on the customer's overall score. From tables 2.2 and 2.3, we have quite a few features related to the customer experience even before they are in the treatment phase. For example, we had about 8 continuous features that tracked the time the interval times between the customer's stages such as time it took for the impression kit to arrive at the customer or the time it took until a treatment plan was created. From our research those times did not have a significant effect on the model. Furthermore, we look at the categorical variable of whether a customer needed to "Rekit" (meaning redo their impressions). This also did not have an impact on their experience despite the inconvenience. Our NPS scores were mainly impacted based on features during the customer's treatment after recieving the aligners. One reasoning is that customer's are hyper focused on the result of their aligners and is easily able to forgive as long as they can see results and factors during treatment are minimized.

6.1.2 Treatment Length Has Both Direct and Indirect Impact

One recurring learning in this research is the direct and indirect effect treatment length has on the score. Recall in section 3.2, that we saw a negative correlation between treatment length and score. That is, for every increase in duration, there is a decline in likelihood of a good score. Also recall that we ran a linear regression between a few feature to predict length and saw a few significant results (refer to table 3.6). While the direct impact to score was not too much of a surprise, this approach of using treatment as an indirect impact compiled from other features is very intriguing and could show that there are signs and reasons we would leave some features that are insignificant in the model.

This is further exemplified when we mention in the prior subsection we mentioned that the features in early stage experience has no direct impact; but we never mentioned that it didn't have any impact at all. It is almost as if treatment length is some form of reduction variable that could explain a bit more of the variation as opposed to adding 5 more features in the model similar to that of a Principal Component Analysis [Far05]. This aspect illustrates the importance of the treatment length in the model.

6.1.3 Cases Data Is Key

One of the critical pieces in the model were the cases data. We saw it in the exploratory data analysis as well as in the model. A couple of reasons why that is. The first is that we saw clear patterns between the number of cases to the score in exploration. This is, the more number of cases opened the lower the mean score. The second is that the cases severity (general vs pro care vs clinical priority). While we saw some puzzling results for the clinical priority, we were able to quantify how much of an impact each cases had on score based on severity. For example, general cases saw a .13 unit decline each additional case but professional care cases saw a .45 unit decline. These features gave us dimensional depth which ultimately proves that as customers go high up in severity there is loss in likelihood

of a promoter score. Further research can be put into this as we can try to solve the clinical priority cases but also look into aspects such as length of the case opened or details on the resolution type and time. We can quite possibly add more interaction variables to help further refine our model.

6.2 Further Research and Improvements

Of course, no research is perfect and there is always room for improvement. Below outlines just a few of the improvements we could make or look into more.

6.2.1 Ordinal Regression Improvement

One of the drawbacks we see in our ordinal regression model was the inability to predict the neutral score. A part of it could be due to the lack of neutral scores the data set has to begin with. Table 6.1, out lines the distribution of the promoter/neutral/detractor scores. Notice that neutral scores represent just 13% of total response. As such, it is likely hard for the model to fine tune towards that's response. A possible way to modify this is by recoding the output so that it would more easily select neutral scores. For example, when compiling the codes of score of "0", we would instead have the cutoff be 40% instead of 50% or above. Another solution would be to add more features or find more outliers to the dataset that would improve model accuracy.

Table 6.1: Breakdown of NPS Scores (Ct. and Pct.)

	Detractor	**Neutral**	**Promoter**
NPS Score Count	25%	13%	62%
% of Total	114	62	286

6.2.2 Supplemental Features

We have proven out that there exists some features that are effective at what NPS scores customers will give. We've also isolated that it's not pre-treatment features that has the more impact, rather it's post treatment. This gives us a better idea of what features we should add in. If we wanted to invest a bit more time in further improving out model, there's a few features we can add in:

1. Average Case Length Time
2. Length of Treatment before requiring Refinements
3. Text Mining Techniques from Cases Data (Sentiment)

The list is based on what we have found to be already be impact to the model and then diving a bit further. Some more exploratory analysis is needed to understand correlation to score but these are just a short list that we should be confident would yield incremental improvements.

6.3 Applications in Real World

NPS Scores is one of the only times companies receive real feedback from consumers. It's a valuable tool to help companies improve on a product or iterate on a new product entirely. Equally important, NPS is a number companies often times tout to customers as it's a direct reflection on the reputation and quality of the company. Because of that, being able to understand what factors into their NPS score as well as being able to predict NPS score can prove to be beneficial to a company as it would allow them to mitigate poor scores when possible and it also allows them to extract general themes of improvement. It's very likely companies already do this today! This all comes down to say that the model we put together is just an example of how important this research topic is and it expands beyond the clear

aligner industry and into all consumer products. Applying statistical analysis to understand consumer satisfaction is a very important concept for companies and it can only grow larger in importance as its potential is limitless!

REFERENCES

[Cha21] Ankit Chauhan. "Random Forest Classifier and its Hyperparameters.", 2021.

[Dav20] Davis David. "Random Forest Classifier Tutorial: How to Use Tree-Based Algorithms for Machine Learning.", 2020.

[Far05] Julian James Faraway. *Linear Models with R.* Chapman Hall/CRC Taylor Francis Group, 2005.

[Far06] Julian James Faraway. *Extending The Linear Models with R.* Chapman Hall, 2006.

[Fou22a] Wikimedia Foundaton. "Bayes' Theorem.", 2022.

[Fou22b] Wikimedia Foundaton. "Logistic regression.", 2022.

[Fou22c] Wikimedia Foundaton. "Random forest.", 2022.

[Fro22] Jim Frost. "Understanding Interaction Effects in Statistics.", 2022.

[Gan18] Rohith Gandhi. "Naive Bayes Classifier.", 2018.

[Gar18] Michael J. Garbade. "Understanding K-means Clustering in Machine Learning.", 2018.

[Gol21] Jeremy Goldman. "What you can learn from disruptive D2C brands.", 2021.

[Hla22] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*, 2022.

[Le18] James Le. "Support Vector Machines in R.", 2018.

[n] OARC Stats. (n.d.). "Ordinal Logistic Regression — R Data Analysis Examples.".

[n22] Wikimedia Commons. (n.d.). "exam pass logistic curve.", 2022.

[Par16] Stephen Parry. "Ordinal Logistic Regression models and Statistical Software: What You Need to Know.", 2016.

[Par21] The Insight Partners. "INDUSTRY NEWS Clear Aligners Market Size to Outstrip $8,708.67Mn by 2028 Growth Projections at 15.9% CAGR During 2021 to 2028 COVID Impact and Global Analysis by TheInsightPartners.com.", 2021.

[Qua22] Qualtrics. "Your guide to net promoter score (NPS) in 2022.", 2022.

[Wis] Fedewa Holder Teichner Wiseman. "Five-star growth: Using online ratings to design better products.".