**Foundations of Data Science**

**Capstone Project: Data Story**

Daniel R. Poston

Leaving money on the table is a problem most companies try to avoid. It’s an issue that presents itself to business in many forms. As a business gains more customers it can become increasingly difficult to determine which of them are likely to order more products, and why. Our focus will be on maximizing revenue from current customers who are likely to make additional purchases.

My client on this project is StandDesk, Inc. They manufacture the most cost effective automatic height adjustable desks on the market ([Receive a $50 discount by ordering here!](http://go.standdesk.co/dT8sl)). After raising over $800,000 on Kickstarter in 2014, they quickly began taking orders from businesses who were trying to create a healthier workspace for their employees. For a lean startup like StandDesk, it can become increasingly difficult to maximize revenue from a rapidly expanding business to business customer base without utilizing a laser like focus of marketing and sales resources on the right customers at the right times. We will attempt to determine which customers to focus on by creating a model that predicts the total revenue that can be generated from each customer.

The data was downloaded from StandDesk’s CRM and Shopify account. The CRM contains information related to interactions between StandDesk and their customers as well as their customers’ demographic information. Data downloaded from Shopify contains information related to purchases from StandDesk’s customers.

The data from the CRM was downloaded as two separate csv files. The first csv file was downloaded from the Companies tab and contains data on all of the companies which have purchased from StandDesk. This data mostly consists of information about each company such as Phone Number, Website, Address, etc., along with demographic data including Employee Count, Industry, Year Founded, etc. The second csv file was downloaded from the Contacts tab which contains data on the contacts related to the companies who have purchased from StandDesk. This data contains information about the contacts such as Phone Number, Email Address, Title, etc. Activity data related to how the contact has interacted with StandDesk is also included in this file. The activity data includes Email communications, calls, website visits, form submissions, etc.

The data from Shopify was also downloaded as two csv files and contains information on each purchase a customer has made. The Order report downloaded from Shopify includes Purchase Date, Purchase Amount, Contact’s Email, etc. Also downloaded from Shopify was the SKU report which contains information on which products were purchased in each order and the related SKUs.

Once each csv file was uploaded to R it was necessary to combine the data sets so that each row represents a different company that has made a purchase from StandDesk. Unnecessary and inaccurate variables were eliminated. The company data frame already contained a field labelled “Domain.” A field labeled “Domain” was created in the Contact, Order and SKU data frames by separating the domain out from the email fields. The data in the Contact, Order and SKU data sets were aggregated based on the domain fields. Each of the four data sets were joined based on the domain fields of each, with rows that did not have a match in the other three data frames being filtered out. Additionally, All NA values were eliminated and companies with less than $1,500 in total revenue were filtered out.

**The following new variables were created:**

**total\_revenue:** The sum of the revenue from all orders related to each of the companies.

**total\_order\_count:** The sum of the number of orders related to each company.

**Days\_Between\_All\_Orders:** The number of days between the first and last order from each company.

**Ave\_Days\_Between\_Orders:** Total days between first and last order divided by Days\_Between\_All\_Orders minus one.

**days\_since\_last\_order:** The number of days since the most recent order for each company.

**days\_since\_first\_order:** The number of days since the first order date for each company and today’s date.

**Ave\_Order\_Amount:** The average amount spent per order for each company.

**Ave\_Reorder:** The average amount spent per order (excluding the first order) for each company.

**after\_cutoff\_date:** A Boolean field where 1 represents companies that made their first order after 2016-01-12, and 0 represents all other companies.

**Total\_Emails\_Delivered:** The sum of emails delivered to contacts related to each company

**Total\_Emails\_Opened:** The sum of emails opened by contacts related to each company.

**Total\_Emails\_Clicked:** The sum of emails clicked by contacts related to each company.

**Emails\_Opened\_Percent:** Emails opened divided by emails delivered.

**Emails\_Clicked\_Percent:** Emails clicked divided by emails delivered.

**Order\_One\_Date – Order\_Fourteen\_Date:** Fourteen date fields for each of the company’s order dates.

**Order\_One\_Amount – Order\_Fourteen\_Amount:** Fourteen fields for the revenue amount of each company’s orders.

**First\_Order\_Traffic\_Source:** The traffic source of each company’s initial order.

**Preliminary Exploration & initial findings**

As mentioned previously, the total lifetime value of a Business to Business customer is what I am attempting to predict. The linear regression model will be created for this purpose using the dependent variable total\_revenue which represents total lifetime value.

StandDesk began using their current CRM on January 13th, 2016. The data related to interactions between contacts and StandDesk is fairly accurate for companies that made a first purchase on January 13th or later. Unfortunately, this is not the case for the data related to companies that purchased prior to January 13th, 2016. This lack of accurate data has become a significant limitation to creating an accurate model. As it stands, the limited data set indicates there is an insufficient number of observations for creating a subset of data for the purpose of testing the model. Therefore, this model will only be useful for predicting total revenue of future customers.

The best linear regression model created so far (see below) has a Multiple R-squared of 0.828 and an Adjusted R-squared of 0.656 from 49 observations. The Residual standard error is 2,625 on 43 degrees of freedom which is fairly high considering the mean total revenue is 3,984.

***lm(total\_revenue ~***

***Ave\_Reorder + days\_since\_last\_order + Industry + Total\_Emails\_Opened\*Total\_Emails\_Delivered, …)***

My goal for finalizing the project will be to increase the number of observations while also increasing the Adjusted R-squared. Due to the fact that I will only be using data for companies who made a first purchase after January 12, 2016, it will be necessary to download additional data that was generated over the last few weeks since my previous download. Additional variables will be created in an effort to further dial in the model. Demographic data that has yet to be tested includes Company Age, Yearly Revenue, whether a company is public or private, and an accurate employee count. The employee count is an important variable to test as this should be directly related to the number of desks a company can purchase from StandDesk. This information will need to be gathered from additional online sources.

The primary takeaway at this point is how much the current results underscore the importance of maintaining accurate data within the company CRM. It may be necessary to determine if additional data exists for companies that purchased prior to January 12, 2016 and populate the CRM with any additional data, although that is beyond the scope of this project.