Predicting Future Revenue: StandDesk B2B Customers

Daniel R. Poston
www.sharpdata.co
July 31, 2016

Overview

- The Client
  - StandDesk
- The Problem
  - Maximizing the lifetime value of B2B Customers

Pre Project Data Prep

- The original files downloaded from Shopify and Hubspot contain proprietary information.
- It was necessary for this information to be removed and replaced with a generic ID.

Pre Project Data Prep: Code

```
Company_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-companies2016-06-24wID.csv", header=TRUE, sep=",", na.strings = "")
cid_df <- data.frame(Company_Report)

# Importing Shopify order data
Order_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/Sales by customer 6.24.csv", header=TRUE, sep=",", na.strings = "")
orid_df <- data.frame(Order_Report)

orid_df <- separate(orid_df, email, c("Email Prefix", "Domain"), sep = "@")
orid_df <- filter(orid_df, total_sales > 0)
orid_df$Domain <- tolower(orid_df$Domain)

cid_df <- mutate(cid_df, Domain = Company.Domain.Name)
cid_df$Company.Domain.Name <- NULL
cid_df$Domain <- tolower(cid_df$Domain)

sub_cid_df <- select(cid_df, Domain, Company.ID)
orid_df <- inner_join(orid_df, sub_cid_df, by = "Domain")

Contact_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-contacts2016-06-24.csv", header=TRUE, sep=",", na.strings = "")
contactid_df <- data.frame(Contact_Report)

contactid_df <- separate(contactid_df, Email, c("Email Prefix", "Domain"), sep = "@")
contactid_df$Domain <- tolower(contactid_df$Domain)
contactid_df <- filter(contactid_df, Domain != "standdesk.co")
contactid_df <- inner_join(contactid_df, sub_cid_df, by = "Domain")
```

Pre Project Data Prep: Code (2)

```
orid_df$name <- NULL
orid_df$`Email Prefix` <- NULL
orid_df$Domain <- NULL
orid_df$company <- NULL

cid_df$Name <- NULL
cid_df$Street.Address <- NULL
cid_df$Website.URL <- NULL
cid_df$Facebook.Company.Page <- NULL
cid_df$Google.Plus.Page <- NULL
cid_df$LinkedIn.Bio <- NULL
cid_df$LinkedIn.Company.Page <- NULL
cid_df$Twitter.Handle <- NULL
cid_df$Domain <- NULL

contactid_df$First.Name <- NULL
contactid_df$Last.Name <- NULL
contactid_df$Company.Name <- NULL
contactid_df$'Email Prefix' <- NULL
contactid_df$Domain <- NULL
contactid_df$Phone.Number <- NULL
contactid_df$Street.Address <- NULL
contactid_df$Billing.Address.Line.1 <- NULL
contactid_df$Shipping.Address.Line.1 <- NULL
contactid_df$ReferrerEmail <- NULL
contactid_df$IP.Address <- NULL
contactid_df$Website.URL <- NULL

write.csv(orid_df, file = "C:/Users/GamingFoSho/Documents/wdR/Sales by customer 6.24wID1.csv", row.names=FALSE)
write.csv(cid_df, file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-companies2016-06-24wIDclean1.csv", row.names=FALSE)
write.csv(contactid_df, file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-contacts2016-06-24wID2.csv", row.names=FALSE)
```

Data Set

- **Order Report: Sales by customer 6.24wID1.csv (Shopify)**
- **Company Report: hubspot-crm-view-companies2016-06-24wIDclean1.csv (Hubspot)**
- **Contact Report: hubspot-crm-view-contacts2016-06-24wID2.csv (Hubspot)**

Data Wrangling: Importing Files

- **Order Report: Sales by customer 6.24wID1.csv (Shopify)**
  - Order_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/Sales by customer 6.24wID1.csv", header=TRUE, sep=",", na.strings = "")
  - or_df <- data.frame(Order_Report)
- **Company Report: hubspot-crm-view-companies2016-06-24wIDclean1.csv (Hubspot)**
  - Company_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-companies2016-06-24wIDclean1.csv", header=TRUE, sep=",", na.strings = "")
  - c_df <- data.frame(Company_Report)
- **Contact Report: hubspot-crm-view-contacts2016-06-24wID2.csv (Hubspot)**
  - Contact_Report <- read.csv(file = "C:/Users/GamingFoSho/Documents/wdR/hubspot-crm-view-contacts2016-06-24wID2.csv", header=TRUE, sep=",", na.strings = "")
  - contact_df <- data.frame(Contact_Report)

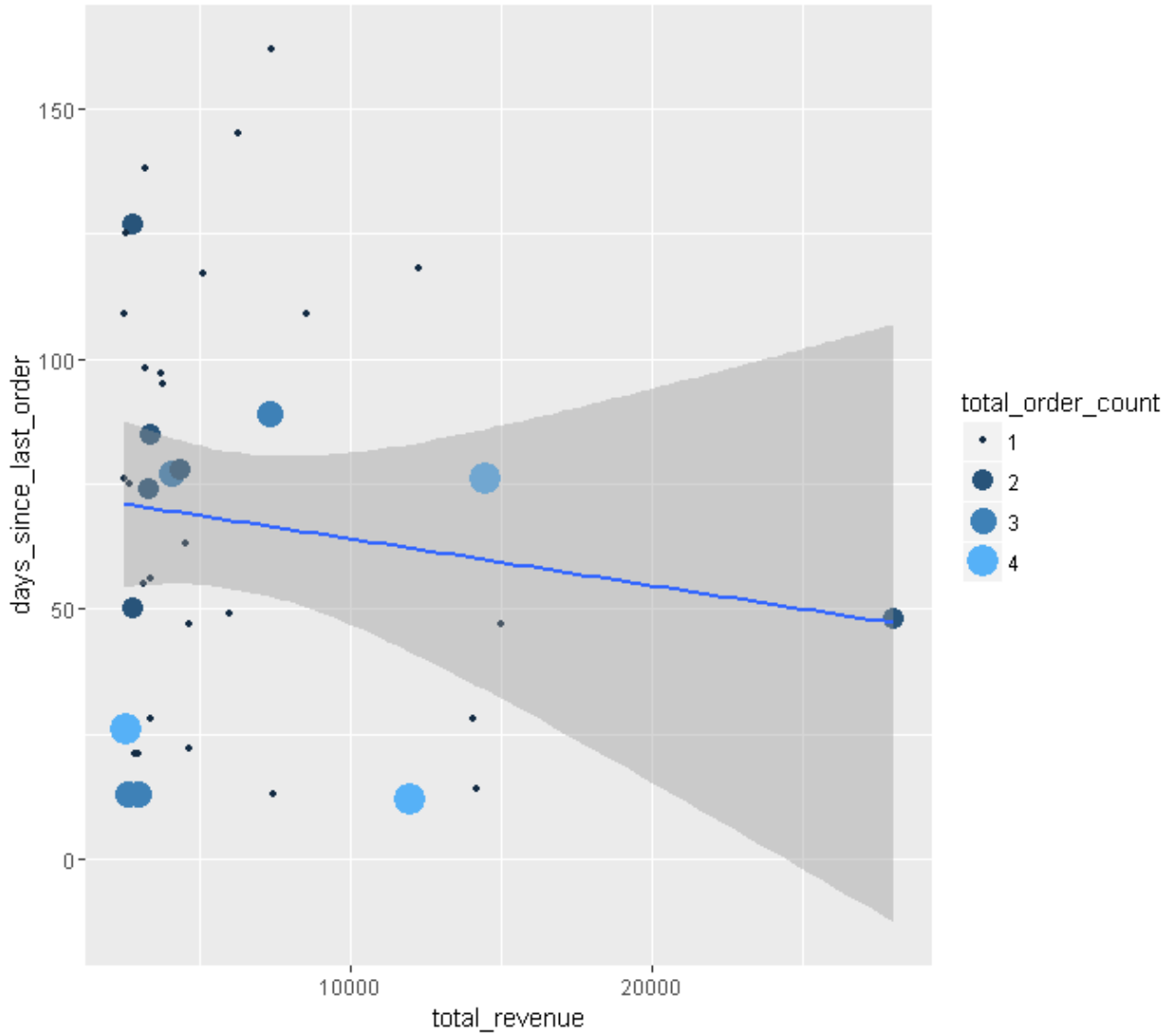
Data Wrangling: New Variables Created

- **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.

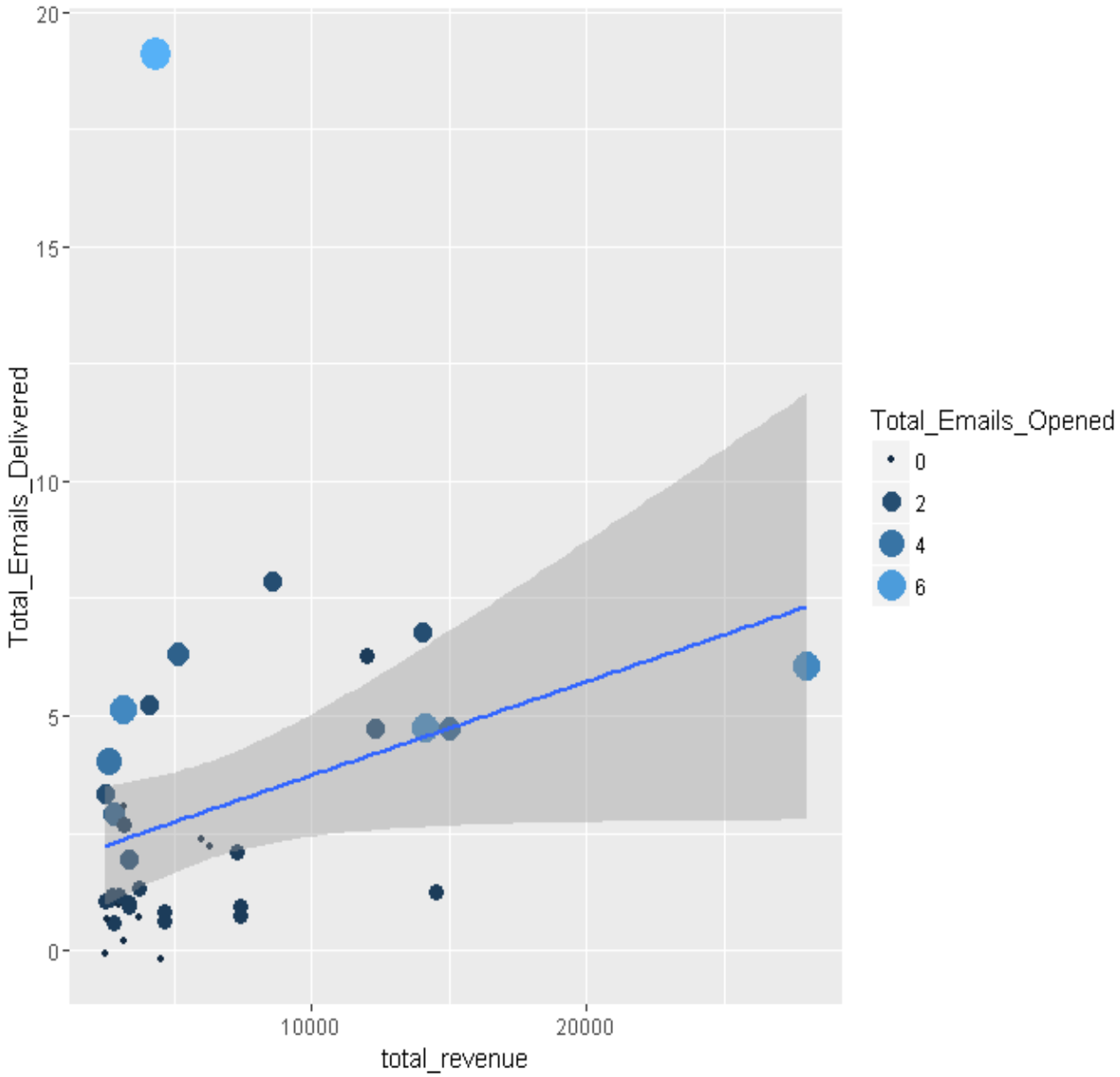
Tidy Data

total_revenue	days_since_last_order	Industry	Total_Emails_Opened	Total_Emails_Delivered
2501.06	109	Unknown	2	3
2514.45	76	Real Estate	0	0
2532.12	125	Unknown	1	1
2541.20	26	Non-Profit Organization Management	0	1
2604.90	13	Non-Profit Organization Management	4	4
2691.45	75	Information Technology and Services	1	1
2778.76	50	Unknown	1	1
2812.51	127	Unknown	1	1
2822.43	21	Unknown	3	3
2965.94	13	Unknown	0	1
2975.64	21	Unknown	1	1
3149.93	55	Sports	5	5
3166.02	14	Unknown	0	0
3166.52	138	Unknown	0	3
3198.22	98	Construction	1	3
3336.14	74	Primary/Secondary Education	1	1
3349.22	28	Oil & Energy	2	2
3369.95	56	Computer Software	1	1
3385.46	85	Unknown	0	1
3744.73	97	Computer Software	1	1
3747.98	95	Telecommunications	0	1
4067.95	77	Machinery	2	5
4367.35	78	Banking	7	19
4556.18	63	Events Services	0	0
4641.43	47	Primary/Secondary Education	1	1
4667.20	22	Accounting	1	1
5119.70	117	Hospital & Health Care	3	6
5989.17	49	Architecture & Planning	0	2
6285.86	145	Information Technology and Services	0	2
7299.51	89	Unknown	1	2
7398.02	162	Unknown	1	1
7418.67	13	International Trade and Development	1	1
8561.58	109	Marketing and Advertising	2	8
11998.07	12	Market Research	1	6
12296.26	118	Unknown	2	5
14054.24	28	Music	2	7
14174.95	14	Renewables & Environment	5	5
14500.30	76	Consumer Goods	1	1
14995.92	47	Semiconductors	3	5
28050.00	48	Unknown	5	6

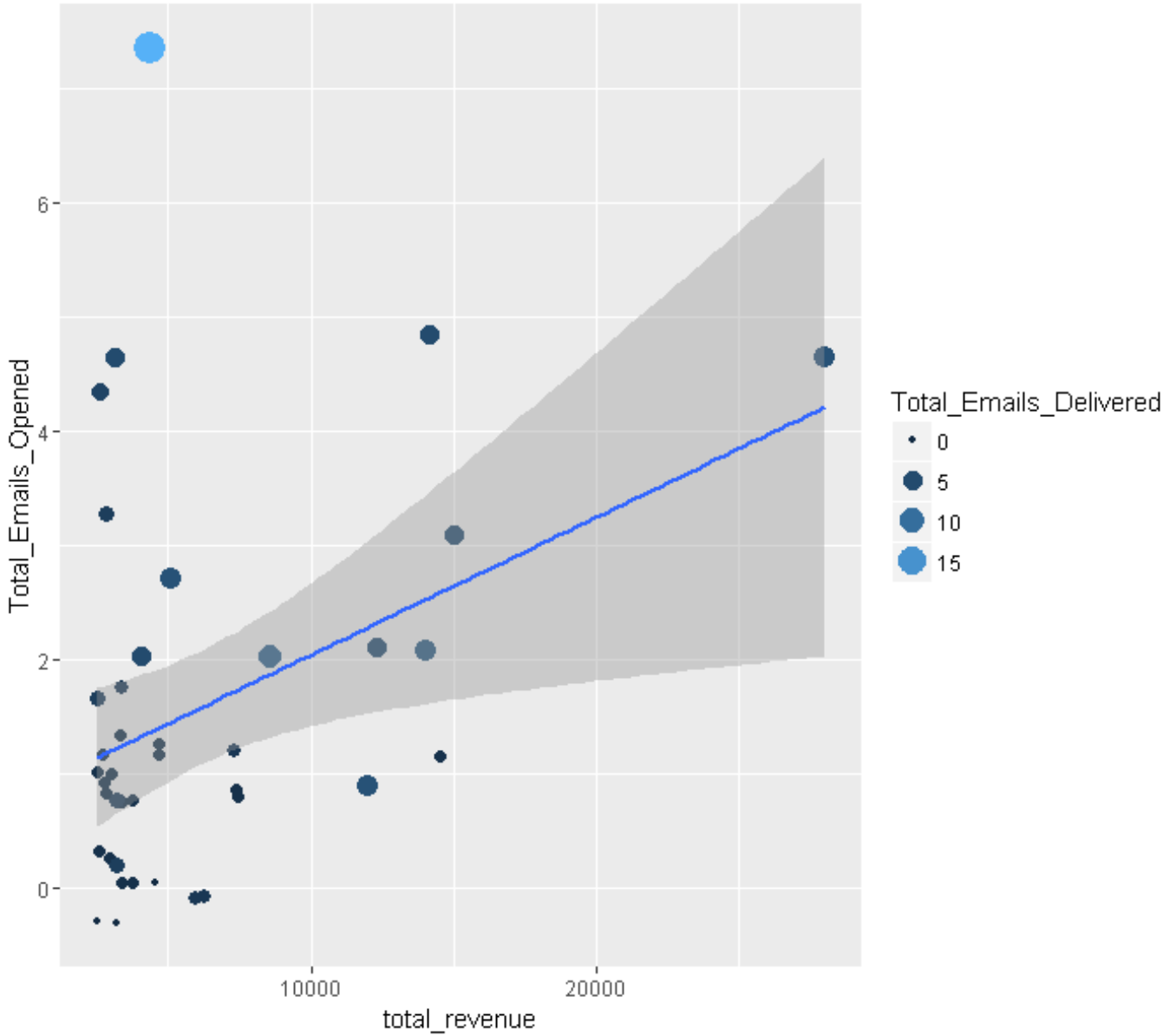
Days Since Last Order Graph



Total Emails Delivered Graph



Total Email Opened Graph



Linear Regression Model

modelsub1 <- lm(total_revenue ~ days_since_last_order + Industry + Total_Emails_Opened*Total_Emails_Delivered, data = sub_company_df)

summary(modelsub1):

	Pr(>|t|)
(Intercept)	0.007163 **
days_since_last_order	0.073802 .
IndustryArchitecture & Planning	0.655837
IndustryBanking	2.31e-06 ***
IndustryComputer Software	0.394870
IndustryConstruction	0.171733
IndustryConsumer Goods	0.025273 *
IndustryEvents Services	0.162865
IndustryHospital & Health Care	0.000893 ***
IndustryInformation Technology and Services	0.256142
IndustryInternational Trade and Development	0.373463
IndustryMachinery	0.033101 *
IndustryMarket Research	0.433913
IndustryMarketing and Advertising	0.019597 *
IndustryMusic	0.585836
IndustryNon-Profit Organization Management	0.018171 *
IndustryOil & Energy	0.804494
IndustryPrimary/Secondary Education	0.561998
IndustryReal Estate	0.053572 .
IndustryRenewables & Environment	0.046556 *
IndustrySemiconductors	0.820650
IndustrySports	0.000140 ***
IndustryTelecommunications	0.112001
IndustryUnknown	0.169794
Total_Emails_Opened	0.000817 ***
Total_Emails_Delivered	0.257120
Total_Emails_Opened:Total_Emails_Delivered	2.32e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2292 on 13 degrees of freedom
Multiple R-squared: 0.9356, Adjusted R-squared: 0.8069
F-statistic: 7.269 on 26 and 13 DF, p-value: 0.0002812

Conclusion

- Unfortunately, the final result of this project is that we currently have insufficient accurate data to use to predict future revenue of StandDesk's B2B customers.
- We have a myriad of ways of collecting and storing data related to potential and current customers and that data is extremely valuable if utilized properly.
- On the other hand, if it is not collected or stored properly, it has no value to us.