

Target Marketing and Demand Forecasting for a Plumbing and Drain Services Company

Submitted towards partial fulfilment of the criteria

for award of PGPBA by GLIEMR

Final report

Dushyant Bhavsar – BAPMAR17013

Niravkumar Machhar - BAPMAR17067

Prasanta Panja - BAPMAR17029

B.P. Kiran - BAPMAR17026

PGP-BABI-Pune-March 2017

Project Mentor: Rajesh Jakhotia

Program Director: Dr. S. Basu

Date: January 8th 2018

Version: 1.1

Status: Draft



# Acknowledgements

At the onset we would like to place on record our sincere gratitude and respect for our mentor Mr. Rajesh Jakhotia. We are highly indebted to him for his patience, support, guidance, encouragement, constant supervision, and valuable inputs that helped us complete the project in time. We acknowledge and are grateful to the program director of the Pune Center, Dr. S. Basu for valuable discussions, insights, support and encouragement throughout the project. We would also like to thank all the instructors of the course at the Pune center for discussions, and for providing valuable inputs at different stages of the project.

We would like to take this opportunity to express our gratitude t0 the team at UST Global for providing the data set, discussions and co-operation without which this Capstone project would not have been possible.

We would also like to thank Ms. Gitanjali Devi, Ms. Sapna Kapoor, Ms. Malathi, and Ms. Ambika for their support in conducting the course in Pune. It was instruction and encouragement from them that kept us on our feet, and enabled us to complete the project on time. Completing this project on time would not have been possible without the kind support and help of many individuals, including the support staff who helped in conducting the course at the Pune center. We would like to extend our sincere thanks to all of them.

Last but not the least, we would like to thank our families for their patience, understanding, encouragement, standing by us, and willingness to sacrifice weekends at this stage of our careers, while we pursued new knowledge.

We certify that the work done us in conceptualizing and completing this project is original and authentic. Any information that has been taken from other sources has been duly acknowledged.

Finally, enrolling in this course helped us gain new knowledge, provided us with a new perspective to look at data, increased our understanding, and broadened our horizons.

Dushyant Bhavsar, Niravkumar Macchar, Prasantha Kumar Panja, and B P Kiran

Date: January 13, 2018

Place: Pune

# Certificate of Completion

I hereby certify that the project titled “Target Marketing and Demand Forecasting for a Plumbing and Drain Services Company” was undertaken and completed under my supervision by Dushyant Bhavsar, Niravkumar Macchar, Prasantha Kumar Panja, and B P Kiran, students of the March 17 batch of the Postgraduate Program in Business Analytics and Business Intelligence (PGPBABI) program in Pune. I certify that the work done by the team in conceptualizing and completing this project is original and authentic. Any information that has been taken from other sources has been duly acknowledged.

Mentor: Rajesh Jakhotia

Date: January 13, 2018

Place: Pune

Table of Contents

[Acknowledgements 1](#_Toc503062406)

[Certificate of Completion 2](#_Toc503062407)

[Table of Figures 5](#_Toc503062408)

[List of Tables 7](#_Toc503062409)

[1.0 Introduction 8](#_Toc503062410)

[1.1 Title and objective of the study 8](#_Toc503062411)

[1.2 Need of the study 8](#_Toc503062412)

[1.3 Company under study, data sources, and scope of study 9](#_Toc503062413)

[1.4 Tools and techniques 11](#_Toc503062414)

[1.5 Limitations of the study 11](#_Toc503062415)

[1.6 Link for Complete Dataset 12](#_Toc503062416)

[2.0 Data Description & Preparation 13](#_Toc503062417)

[2.1 Identification of variables 13](#_Toc503062418)

[2.2 Other Useful information and features 21](#_Toc503062419)

[2.3 Data Dictionary 37](#_Toc503062420)

[2.4 Data Quality Concerns and Preparation 39](#_Toc503062421)

[2.5 Data Preparation 39](#_Toc503062422)

[3.0 Hypothesis Testing 42](#_Toc503062423)

[3.1 List of Null hypotheses: 42](#_Toc503062424)

[3.2 Correlation matrix 43](#_Toc503062425)

[3.3 Hypothesis testing using linear regression equation 44](#_Toc503062426)

[3.4 List of Significant Variables for Model Building 44](#_Toc503062427)

[4.0 Modelling 45](#_Toc503062428)

[4.1 Proof of concept(POC) 45](#_Toc503062429)

[4.2 Logistic Regression 46](#_Toc503062430)

[4.3 Random Forest 46](#_Toc503062431)

[4.4 Decision Tree 47](#_Toc503062432)

[4.5 Naive Bayes 47](#_Toc503062433)

[4.6 Final model using R Studio 48](#_Toc503062434)

[4.7 Analysing RF model on unseen holdout data 51](#_Toc503062435)

[5.0 Conclusions and Recommendations 52](#_Toc503062436)

[5.1 Conclusions 52](#_Toc503062437)

[5.2 Recommendations 52](#_Toc503062438)

[Appendix 53](#_Toc503062439)

[References 53](#_Toc503062440)

# Table of Figures

[**Figure 1: Snapshot of data provided by UST Global** 10](#_Toc503062441)

[**Figure 2: Geographical presence of the plumbing company** 13](#_Toc503062442)

[**Figure 3: Top 5 branches for revenue and number of service calls** 14](#_Toc503062443)

[**Figure 4: Service calls from Jan 2011 to Dec 2013 on scale of % # of Calls and % of Revenue** 17](#_Toc503062444)

[**Figure 5: Service call throughout week** 18](#_Toc503062445)

[**Figure 6: Service call throughout Day** 19](#_Toc503062446)

[**Figure 7: Sales drilled down to transaction level** 20](#_Toc503062447)

[**Figure 8: Total calls month wise 2011 (1=Jan, 12= Dec)** 21](#_Toc503062448)

[**Figure 9: Total calls month-wise 2012 (1=Jan; 12=Dec)** 22](#_Toc503062449)

[**Figure 10: Total calls month wise 2013 (1=Jan, 12= Dec)** 22](#_Toc503062450)

[**Figure 11: Business from Customer Type across geography** 23](#_Toc503062451)

[**Figure 12: Top revenue generating job codes** 25](#_Toc503062452)

[**Figure 13: Average time (hours) taken to complete the job** 26](#_Toc503062453)

[**Figure 14: Churn data for latency (days)** 27](#_Toc503062454)

[**Figure 15: Churn by customer’s count of distinct job codes** 28](#_Toc503062455)

[**Figure 16: Churn by number of calls from customer till date** 29](#_Toc503062456)

[**Figure 17: Churn by number of days elapsed since last call date** 30](#_Toc503062457)

[**Figure 18: Churn by sum of service delay in(hrs.)** 31](#_Toc503062458)

[**Figure 19: Churn by sum all service ticket revenue for individual customer** 32](#_Toc503062459)

[**Figure 20: Churn by customer lifetime** 33](#_Toc503062460)

[**Figure 21: Churn data by customer type** 34](#_Toc503062461)

[**Figure 22: Churn data by customer type on day basis** 35](#_Toc503062462)

[**Figure 23: Percentage of yearly new versus repeat customers** 36](#_Toc503062463)

[**Figure 24: Correlation matrix** 43](#_Toc503062464)

[**Figure 25: Output for correlation matrix** 43](#_Toc503062465)

[**Figure 26: Knime output showing model building approach** 45](#_Toc503062466)

[**Figure 27: Confusion matrix of the Logistic Regression Model** 46](#_Toc503062467)

[**Figure 28: Confusion matrix of the Random Forest Model** 46](#_Toc503062468)

[**Figure 29: Confusion matrix of the Random Forest Model** 47](#_Toc503062469)

[**Figure 30: Confusion matrix of the Naïve Bayes Model** 47](#_Toc503062470)

[**Figure 31: Performance output of final RF model (development model)** 48](#_Toc503062471)

[**Figure 32: OOB error rates of Random Forest development model** 49](#_Toc503062472)

[**Figure 33: mTry curve for finding optimum** 49](#_Toc503062473)

[**Figure 34: AUC Curve of Random Forest development mode** 50](#_Toc503062474)

[**Figure 35: Performance output of final RF model (holdout data)** 51](#_Toc503062475)

# List of Tables

[**Table 1: Tools and techniques** 11](#_Toc503061119)

[**Table 2: Branch ID by State, number of records and total revenue** 15](#_Toc503061120)

[**Table 3: List of variables that define consumer behavior** 16](#_Toc503061121)

[**Table 4: Revenue from base customers classified by type** 23](#_Toc503061122)

[**Table 5: Revenue and ticket counts of Job Codes with revenue greater than $ 50,000** 24](#_Toc503061123)

[**Table 6: Explanation of variables present in the data set** 37](#_Toc503061124)

[**Table 7: List of Null Hypothesis statements** 42](#_Toc503061125)

[**Table 8: Statistics on Linear Regression** 44](#_Toc503061126)

[**Table 9: Significant variables for model building** 44](#_Toc503061127)

# 1.0 Introduction

## 1.1 Title and objective of the study

The title of the study is “Target Marketing and Demand Forecasting for a Plumbing and Drain Services Company” in which data for the company for a three-year period between January 2011 and December 2013 was analyzed. The objective of the study was to look at the data, derive insights, and based on these insights, provide a recommendation on what is required to make such companies viable, remain competitive, and be successful in the market place. Another intended goal of the study is to provide anyone wishing to enter this area with a realistic expectation of the nature of the business, the amount of business they can expect, they revenues they can expect, and what can be done to attract new customers, and more importantly retain old customers.

## 1.2 Need of the study

Most of us must have personally experienced not able to leave home either to go to work or for running chores because of an overflowing pipe or a tap that does not seem to close. It would have been a distressing experience which would left the home owner, for lack of a better word, drained, and the house at an elevated state of mess. At such times, Murphy’s law gains precedence, and the search for the elusive plumber becomes even more difficult. This is not because there are no plumbers in the vicinity or that the plumbers are not available, it is because we just do not know how or where to look for them. The most obvious place to search is the local yellow pages, but after that, it is just word of mouth. In fact, till the recent past, even the plumbers relied on word of mouth for business growth. In such a situation, imagine if there was a phone/internet/mobile app based plumbing services company that offered demand based services to customers at a time suited to the customers. It is not difficult to imagine that the plumbing services company would have attracted a bunch of loyal customers provided they carried out the task well, at a reasonable cost, and within a reasonable time frame.

In the United States, the plumbing and the drain services market is a $ 105 billion market which, between 2011-2016, has grown at an annual average rate of 4%, and is expected to grow at a higher rate over the coming years. In 2014, this industry had about 120,000 businesses, employing over 470,000 employees [[[1]](#endnote-1) ]. Several startup companies like Housecall, Austin Plumbers, Abey Plumbers Bible, Plumbers Order Book, and Plumbers in Witney are looking to make hay while the sun on the on-demand plumbing market shines by adopting a suitable on-demand solution. Companies like Jaggernaut and Hyperlink Infosystem [[[2]](#endnote-2),[[3]](#endnote-3) ] offer customized game-plans and playbooks to companies wishing to transform the on-demand plumbing services sector. In India, the plumbing industry comes under the home services industry which is currently a highly unorganized segment. It is also estimated to be a $100 billion market, and has seen rapid growth over the past few years. At present, it is the hottest area for startup investments in which technology is causing disruptions. In India, several companies including Urban Clap [[[4]](#endnote-4)], and Gapoon [[[5]](#endnote-5)] are also providing mobile app based on demand plumbing services. The objective of most of these startups is to bring standardization into this highly unorganized market in India. In 2014 alone, 69 startups were founded in the homes services industry. It is no wonder that the sector is witnessing an influx of funds from angel investors and institutional investors alike. Aditya Rao, CEO of LocalOye expects the home service industry to mirror the current activity shown in the e-commerce segment in India with 1 or 2 players emerging as clear market leaders” [[[6]](#endnote-6) ].

In summary, the timing of this study could not have been better. Insights obtained by analyzing the data would go a long way in helping readers understand the nature and type of work of these businesses, the problems that such businesses face, and a realistic picture on the expectations on the amount of work and the revenues generated by such businesses. An attempt would also be made to provide recommendations on what additional services that business can offer for it to remain competitive and sustain in the market place.

## 1.3 Company under study, data sources, and scope of study

The data has been provided by UST Global, a US multinational company providing Digital, IT services and solutions. The company specializes in Healthcare, Retail & Consumer Goods, Banking & Financial Services, Telecom, Media & Technology, Insurance, Transportation & Logistics and Manufacturing & Utilities. The data belongs to an on-demand plumbing services company that is based in the US. The name of the company, customer information, job types, and types of service provided by the company has not been provided by UST Global. The data contains information for a three-year period between January 2011 and December 2013. The data comprises of 5284 rows and 31 columns (variables), the snapshot of which is shown in Figure 1. The complete data set can be accessed by clicking the icon in the section titled, “Link for Complete Dataset”. The main variables include customer ID, customer type, customer name, address, and contact information, account setup date, last service date, revenue earned, date and time of call for service, and date and time the task was completed. In short information on the customer, type of service required by the customer, date and time the services were requested, date and time the service was completed, revenue earned from the service is provided, and date of the most recent service availed by the customer.



**Figure 1: Snapshot of data provided by UST Global**

The scope of the project is to study the customer's data and to predict the number of days when the customer is likely to call next and for which service type. The project will be executed in four phases. Data preparation will be carried out during the first phase of the project. In this phase, data will be cleaned, if required, missing value will be added. The trends for each variable would be studied, and the most important variables will be identified. During the second phase, the associated generic models for holistically solving the problem of customer churn will be studied. The accuracy of the models developed using Machine Learning (ML) tools will be measured, and finally the directions for further development will be assessed. The third phase will involve development of strategy and game plan for Target Marketing. A comprehensive strategy will involve opportunity identification (identification of opportunities to attract), convert, close, and please each customer. The strategy starts with goals, personas, keyword research, competitive analysis, content creation, link building, social media marketing, and email marketing. The fourth and final phase will consist of developing a model for demand forecasting.

## 1.4 Tools and techniques

The tools and techniques used in this study are listed in Table 2.

**Table 1: Tools and techniques**

|  |  |  |
| --- | --- | --- |
| S. No | Task | Tools |
| 1 | Data visualization, hypothesis forming | Excel, Tableau, R-Studio, Knime, |
| 2 | Modeling | R Studio (Random Forest, Logistic Regression, Decision Tree, and Naïve Bayes) |
| 3 | Prediction | Simple mathematical calculations |

## 1.5 Limitations of the study

There are several limitations in the study. The most significant was the quality of data provided by UST Global. The data was not very rich, data secrecy laws prevented UST Global from giving out customer information. No meta data was provided on what the data in the columns represented. It was left to the analyst to define what the data in the columns represented and develop an understanding based on this. There was also no definition on the type and nature of services offered by the company or requested by the customer. Based on the data available we could at best develop churn models to predict customer churn. The volume and quality of data provided did enable accurate prediction of customer demand, however a reasonable prediction based on simple mathematical calculations could be made. The original mission and objective of the project, subject to data quality was as given below.

1.5a Intended Mission

The mission of Plumbing & Drain Service company (PDS) is to offer a comprehensive and finest array of new plumbing installations, repairs, sewer, and drain cleaning services using advanced technologies. The main goal of the business is to study, attract, and retain customers by addressing the challenging problem of demand forecasting and customer churn. This can be done by collecting and analyzing historical data of customers along with the kind of services offered to them, to predict/forecast when the customer is likely to call next and for which type of service. Analyzing the churn data shall enable the company to do target market along with segmentation and do forecasting and be better equipped to handle requirements.

1.5b Intended Objectives

The key objective is to position PDS as a premier residential/commercial plumbing service provider in the given area. The marketing strategy will initially generate awareness among the target customers, followed by development of a customer base, and then work towards the development of long-term customers. The message company seeks to communicate is the offer of a high-quality work product supported by industry benchmarked professionalism and customer service. The message will be communicated through two methods. The first is a comprehensive networking campaign that leverages the personal and professional contacts the company has developed during their period of service within the industry. The message can also be communicated by an advertising campaign.

1.5c Intended Marketing Objectives

1. Increase repeat customers per quarter
2. Ensure 100% satisfaction with every job
3. Decrease customer acquisition costs per year

1.5d Intended Financial Objectives

1. Visible profitability within a year
2. Target a double-digit growth rate for the first three years
3. Use data mining and analytics to become a premier customer centric plumbing service provider from just being a plumbing service provider

## 1.6 Link for Complete Dataset

Please click on icon below to access the complete dataset.



# 2.0 Data Description & Preparation

## 2.1 Identification of variables

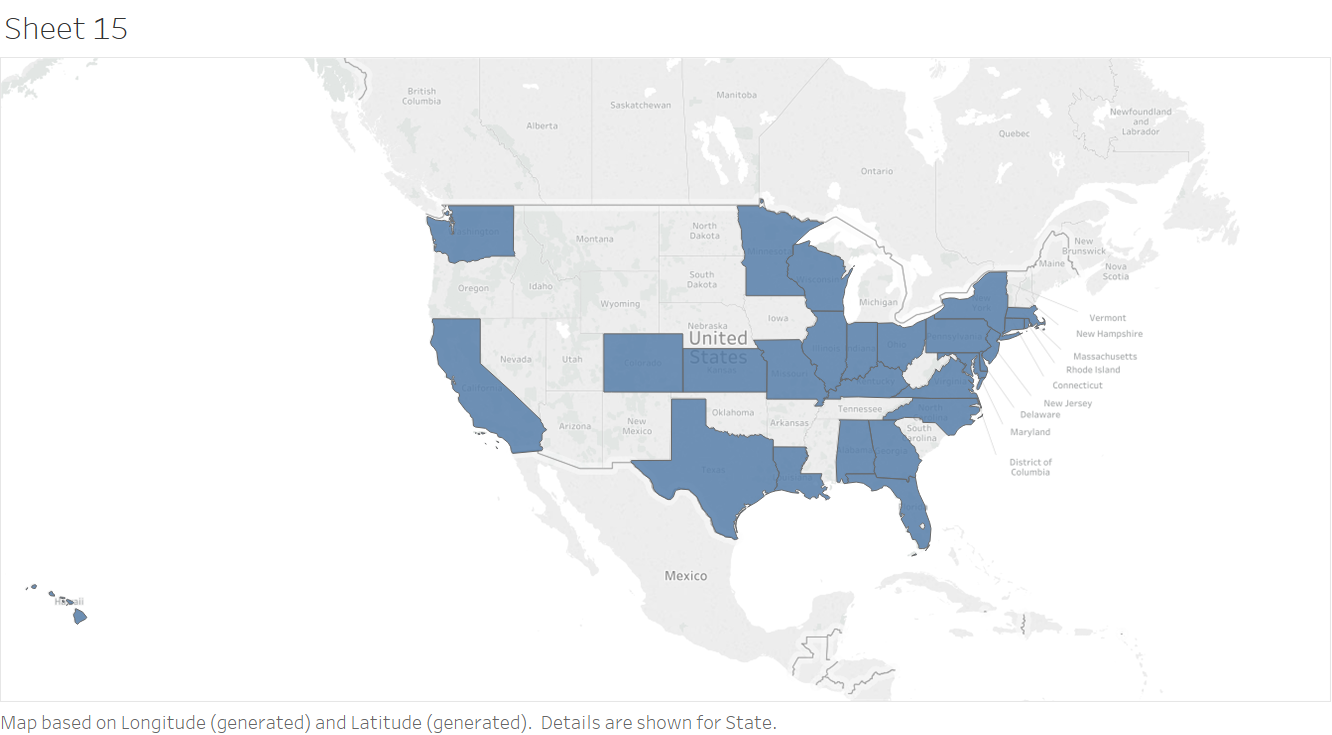
After observing the master dataset and understanding all the metrics and dimensions that were captured by PDS, following variables were extracted:

***Demographics and Geography:***

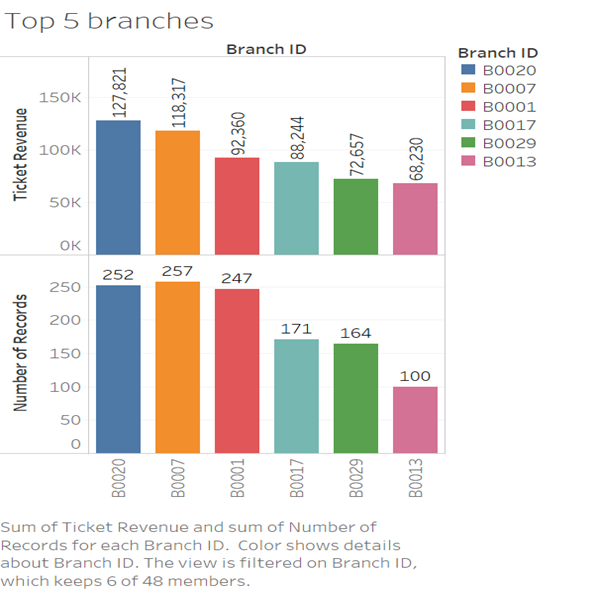
In this category, Unique Customer ID is assigned to every new customer. Data also has few other demographic and geography variables such as Customer Name, Address, City, State, Zip Code, Area, Branch ID etc. Since this is live data from client most client details except State/City is masked and hence dropped for analysis.

***Insights:***

Most company branches have 1:1 mapping with state, whereas few states have been covered by multiple branch depending on state geography spread.

****

**Figure 2: Geographical presence of the plumbing company**



**Figure 3: Top 5 branches for revenue and number of service calls**

The company operates a total of 47 branches across 28 states of the United States of America. Each branch mostly covers a single state; however, a few branches cover multiple states, and a few states have multiple branches. The geographical presence is shown in Figure 2, and details of the Branch ID, the state in which its operates, the number of ticket counts and the sum of ticket revenue for each of these branches is given in Table 2. Overall the company has received 5069 service calls and has generated a revenue of $2,274, 460 within the three years. There are 4 branches with a cumulative revenue of over $80,000 (Branches 20, 7, 1 and 17) over the three years, and 37 branches with revenue more than $ 50,000 over the same time frame. The top 5 revenue generating branches, and the branches that receive the most number of service calls are shown in Figure 3.

**Table 2: Branch ID by State, number of records and total revenue**

|  |  |  |  |
| --- | --- | --- | --- |
| Row Labels | Count of State | Sum of Sum of Ticket Revenue | Sum of Count of Ticket Number |
| B0001 | **6** | **92360.03** | **247** |
| 2011 | 2 | 25692.82 | 77 |
| 2012 | 2 | 33458.1 | 90 |
| 2013 | 2 | 33209.11 | 80 |
| B0007 | **9** | **118316.94** | **257** |
| 2011 | 3 | 24507.8 | 65 |
| 2012 | 3 | 38438.68 | 88 |
| 2013 | 3 | 55370.46 | 104 |
| B0017 | **3** | **88243.76** | **171** |
| 2011 | 1 | 15242.35 | 46 |
| 2012 | 1 | 24960.1 | 50 |
| 2013 | 1 | 48041.31 | 75 |
| B0020 | **3** | **127820.77** | **252** |
| 2011 | 1 | 41672.55 | 75 |
| 2012 | 1 | 38749.97 | 87 |
| 2013 | 1 | 47398.25 | 90 |
| B0029 | **3** | **72656.66** | **164** |
| 2011 | 1 | 20475.99 | 37 |
| 2012 | 1 | 24442.17 | 70 |
| 2013 | 1 | 27738.5 | 57 |
| Grand Total | **24** | **499398.16** | **1091** |

***Acquisition:***

How the customer is acquired, here for this plumbing business, all customers scheduled their appointments via making a call to branch/contacts. No other details on different acquisition medium were provided by company for analysis.

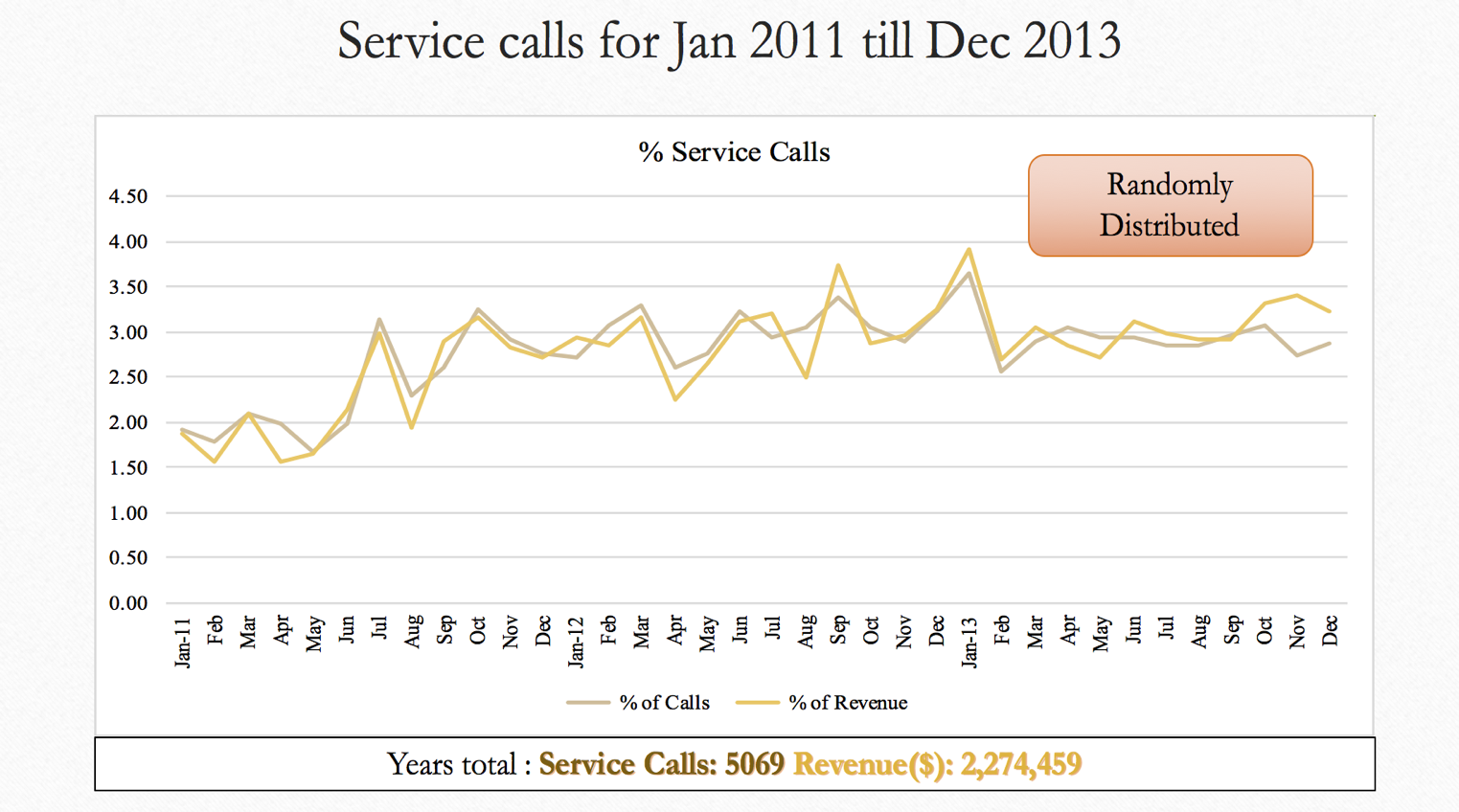
***Behavior:***

This is very significant category for the study. Consumer Behavior or the Buyer Behavior is referred to the behavior that is displayed by the individual while they are, consuming any product or services. These behaviors can be affected by multiple factors and help us meet customer demand more efficiently. This also help understand why, when, how, what and other factors that influence buying decision of the consumers. This category comprises of variables given in Table 3.

**Table 3: List of variables that define consumer behavior**

|  |  |
| --- | --- |
| Feature | Description |
| Customer ID | Unique ID for each customer |
| Cust Life Time | Years given customer used service |
| Last Service Date | Date when last service request complete for customer |
| Last Service Month | Month when last service request complete for customer |
| ShowIn2011 | Binary variable to flag customer shown in 2011 |
| ShowIn2012 | Binary variable to flag customer shown in 2012 |
| CntOfDistinctJobs | No. of different service Jobs used by customer |
| Customer Type | Given customer Id belongs to which type: Type1, type2 or type3 |
| DaysSinceLastCall | This should be calculated runtime when we decide to run model, Number difference between (Today's date - last service date), here we considered Todays date as 31-12-12(Train), 31-12-13(Test) |
| ServiceDelayInHrs | Sum of total service delay in hours for customer (Dispatch time minus Schedule time) |
| AvgHrsToCompleteJob | Average hours taken to complete job for given customer |
| AvgLatencyInDays | Average number of days between calls for given customer |
| IsCalledOnMon | Binary variable to flag customer called on Monday |
| IsCalledOnTue | Binary variable to flag customer called on Tuesday |
| IsCalledOnWed | Binary variable to flag customer called on Wednesday |
| IsCalledOnThu | Binary variable to flag customer called on Thursday |
| IsCalledOnFri | Binary variable to flag customer called on Friday |
| IsCalledOnSat | Binary variable to flag customer called on Saturday |
| IsCalledOnSun | Binary variable to flag customer called on Sunday |
| Sum of Ticket Revenue | Sum of total revenue generatged from given customer |
| Count of Call Date | Number of calls made by given customer |
| Churn | whether customer has churned or not: 1 - churned, 0 - Returning |

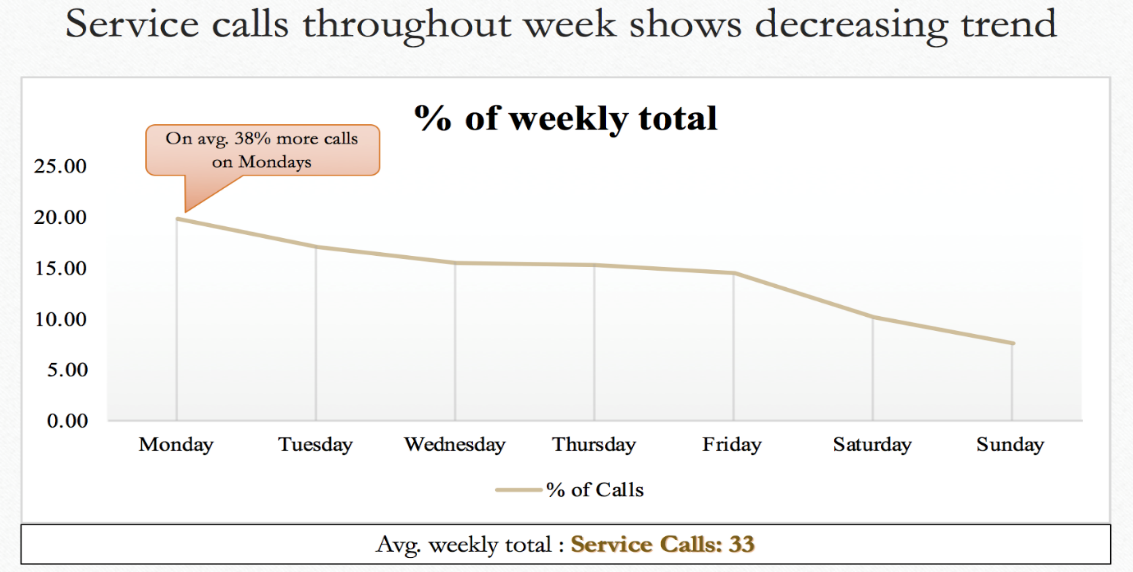
***Trend and seasonality for service calls (Jan 2011 to Dec 2013):***



**Figure 4: Service calls from Jan 2011 to Dec 2013 on scale of % # of Calls and % of Revenue**

Figure 4 shows trend of service calls over between Jan 2011 to December 2013. A total of 5069 service calls were made that generated a revenue of $ 2274459 for the company. Overall, both curves are randomly distributed with no discernable trend visible. This can be attributed to the random nature of the plumbing business as customers call only when there is a requirement. There is generally no fixed planned or routine maintenance activities for these services. Service providers can come up with innovative ideas that attract customers for routine service and maintenance activities in this area.

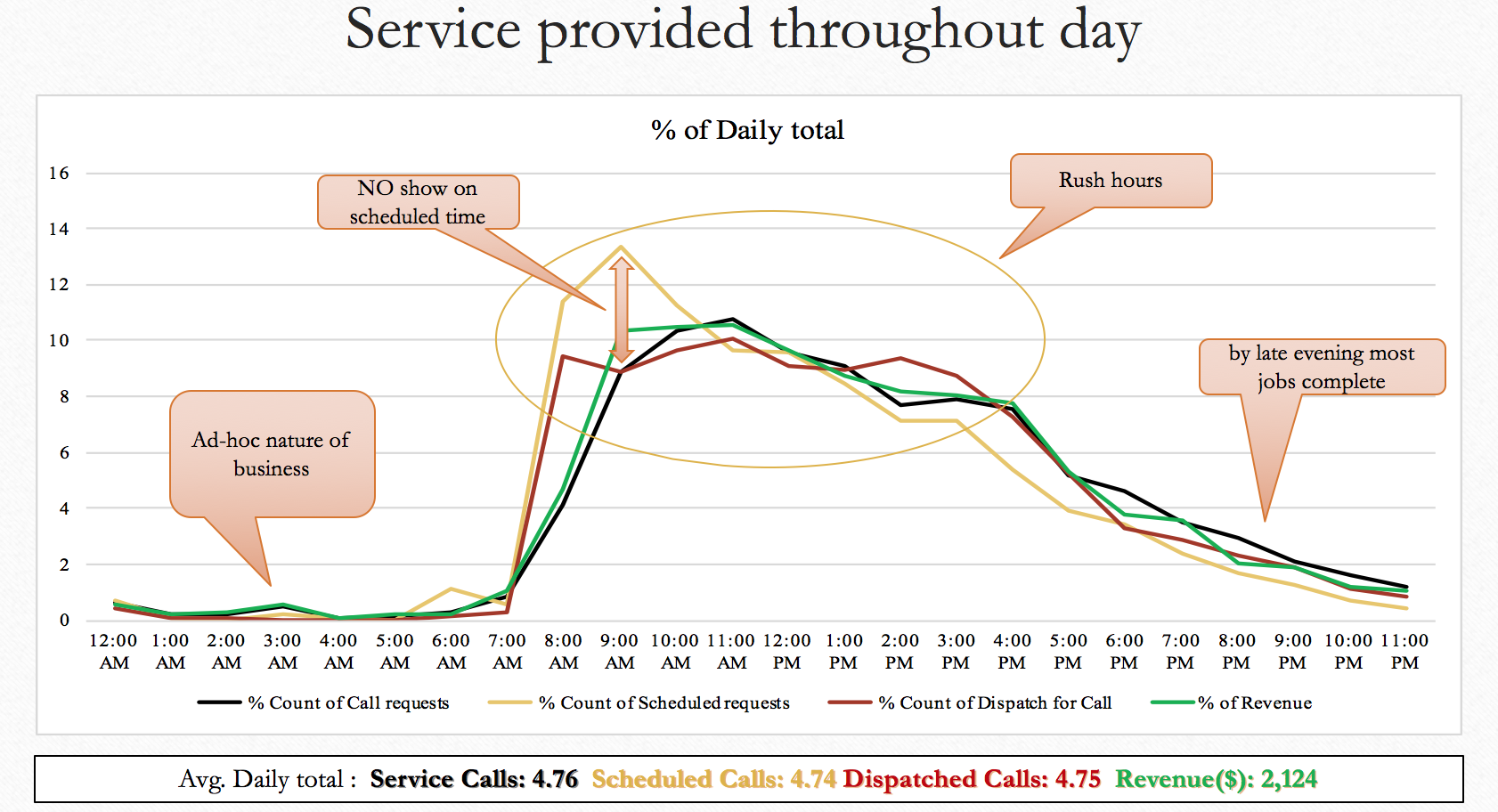
***Day of week analysis:***

.

**Figure 5: Service call throughout week**

A typical week of the business is shown in Figure 5, with Mondays receiving on an average 38% more calls compared to the rest of the week. On an average around 33 service calls are received per week across all branches. In general, the number of service calls falls from Monday to Wednesday. This can be attributed to the general home owner/customer tendency of putting away routine maintenance tasks for the weekend as they get busy with the work week. Thursdays and Fridays see a slight increase in the number of calls. Saturday and Sunday see the lowest number of calls, and these could mainly be attributed to emergency calls

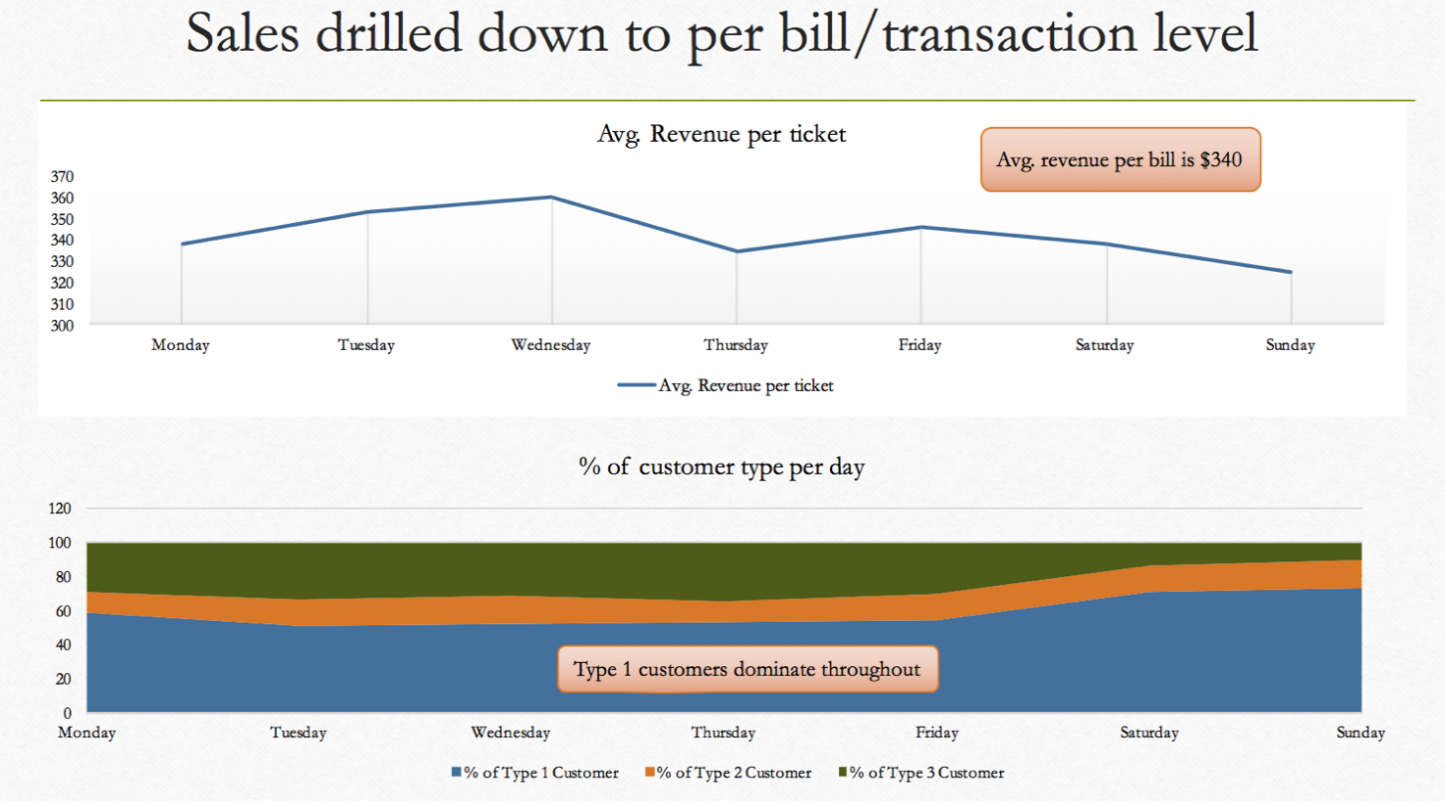
***Hour of day analysis:***



**Figure 6: Service call throughout Day**

A typical work day is shown in Figure 6. On an average about 4.7 calls are received per day and average revenue earned per day is $ 2124. Generally, due to the ad-hoc nature of the business, service calls are received throughout the day. A very small amount of service calls with corresponding to lower revenues are seen during the early morning hours between 12:00 AM to 7:00 AM. The activities start to increase from 7:00 AM with a peak at 11:00 AM and then begin to decrease towards the end of the day. The peak hours for calls received, plumbers dispatched on service calls, and revenue generated is between 7:00 AM to about 4:00 PM. The service call volume and revenues decrease from 4:00 PM to 11:00 PM.

***Average sales per transaction:***



**Figure 7: Sales drilled down to transaction level**

Data of sales drilled down to transaction level is shown in Figure 7. The average revenue per bill is $ 340. The highest average per transaction is on Wednesday (average $360). The figure also shows that Type 1 customers dominate throughout the week.

***Outcome:***

Churn is our target or response variable. This variable tells us whether given customer has churned or not.

***Our definition of Churn:***

Here, churn can be translated into the following rules:

1. The Churn variable (i.e. the one we want to model) is built by looking back over a 6month period from a reference date, and flagging the customers with a “0” if they used our service on this time frame and “1” otherwise i.e. churned.
2. We have restricted our data to the customers who made more than one service calls from start date, and total purchases at least 30 in value. This rule is to make sure we only look at the high value customers.

## 2.2 Other Useful information and features

***Customer ID:***

There are total 1058 unique customers, the average number of calls per customer is 5, and the average spending per customer over the time frame is $ 2000. The number of calls made by month for 2011, 2012, and 2013 are shown in Figures 8, 9 and 10 respectively. There is no specific trend observed, and this can be attributed to the nature of the business. Approximately 34% of the customers have used the services of the company for in all three years and are considered returning or base customers. They contribute about 30 % of the total revenue generated by the company. On an average about 250 calls are made by new customers every year. Over the same time frame, about 8% (85/1058) of the customers did not choose to use the services of the company after 2012. These are considered as lost customers.

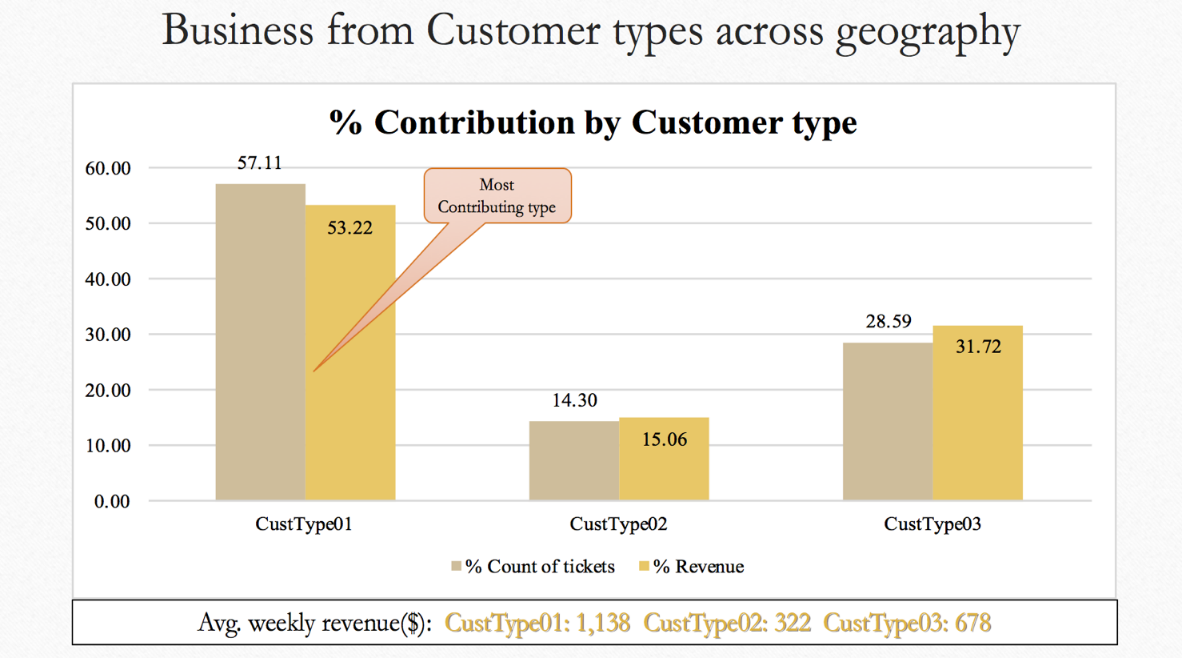
**Figure 8: Total calls month wise 2011 (1=Jan, 12= Dec)**

**Figure 9: Total calls month-wise 2012 (1=Jan; 12=Dec)**

**Figure 10: Total calls month wise 2013 (1=Jan, 12= Dec)**

***Customer Type:***

There are total of three customer types classified as Type 1, Type 2 and Type 3 respectively. Figure 11 shows business by customer type across the entire country. The major contribution in terms of both the number of calls and the revenue generated is from customer Type 1 followed by Type 3. Approximately 57% of the company’s revenue are from Type 1 customers. The average weekly revenue generated by such customers is approximately $ 1138.



**Figure 11: Business from Customer Type across geography**

**Table 4: Revenue from base customers classified by type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer Type | % of Base Customers | Avg. Spend | % Of Calls made | No. of Customers |
| Type 1 | 36.81 | 418.50 | 56.79 | 600 |
| Type 2 | 39.78 | 473.07 | 14.21 | 151 |
| Type 3 | 24.67 | 500.50 | 28.99 | 307 |

Table 4 shows data for base customers classified by type. Again around 57 % of the calls were made by Type 1 customers who make up approximately 37% of the base customers, and spend on an average $ 419. Type 2 customers represent about 40% of base customers and spend on an average $ 473. Type 3 customers make up 25% of the base customers, make around 29% of service calls and spend an average of $ 500. It can be derived that Type 3 customers require specialized tasks that generate higher revenues for the company. More insights can be obtained by analyzing the Job Codes against Customer Type.

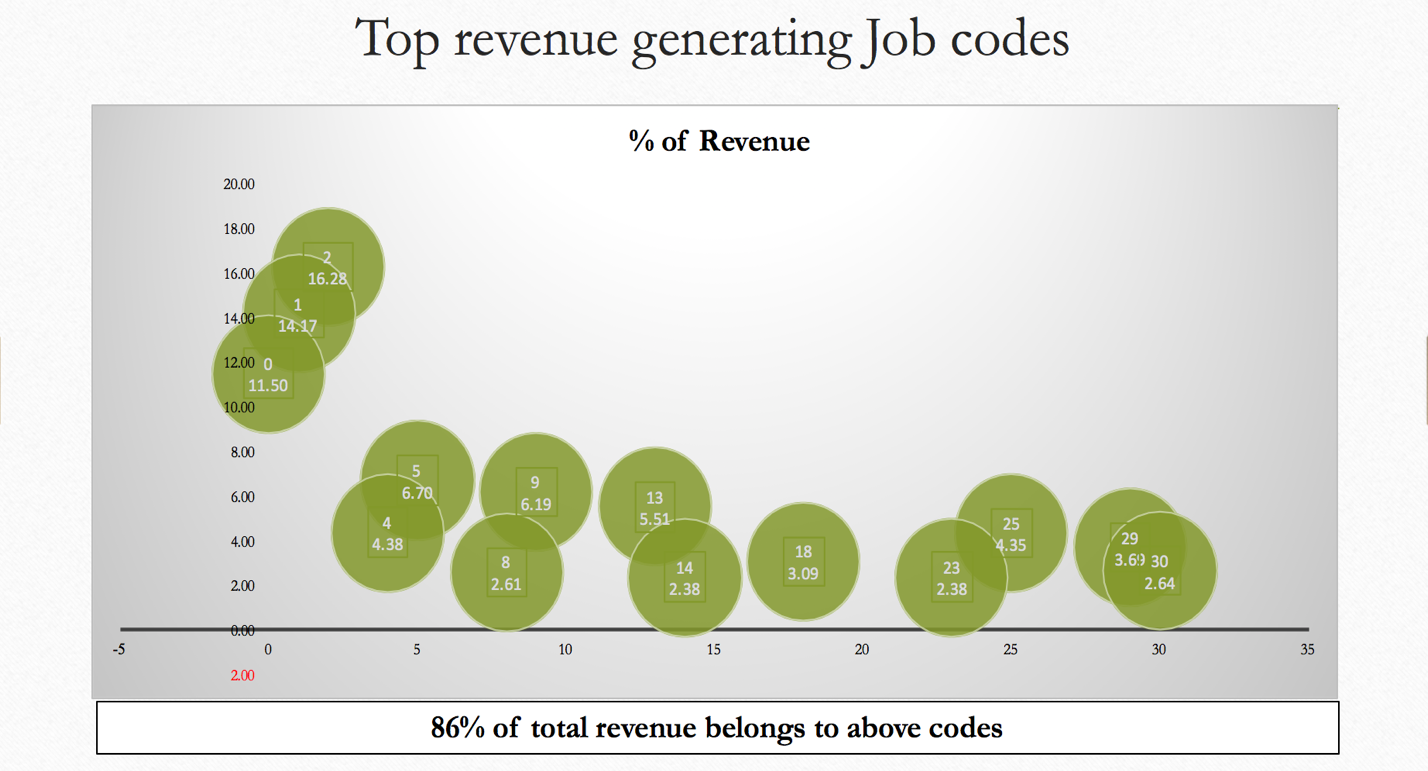
***Job Code:***

Analysis by job code shows that there are 34 unique job codes used across various branches. Table 5 lists data of job codes that are not only used most often, but also generate revenue of greater than $ 50,000 for the period under consideration.

**Table 5: Revenue and ticket counts of Job Codes with revenue greater than $ 50,000**

|  |  |  |
| --- | --- | --- |
| Job Code | Count of Ticket Number | Sum of Ticket Revenue |
| 2 | 631 | 370439.56 |
| 1 | 872 | 322487.37 |
| 0 | 391 | 260756.11 |
| 5 | 622 | 152724.67 |
| 9 | 244 | 141072.03 |
| 13 | 164 | 125651.17 |
| 4 | 300 | 99793.90 |
| 25 | 310 | 99225.37 |
| 29 | 271 | 81517.30 |
| 18 | 45 | 70330.66 |
| 30 | 118 | 60261.26 |
| 8 | 58 | 59455.28 |
| 14 | 90 | 54216.57 |
| 23 | 114 | 54187.72 |
| Grand Total | **4230** | **1952118.97** |

Figure 12 shows data of the top revenue generating jobs codes. The data represented contributes to about 86% of the total revenue generated by the company of which Job codes 2, 1 and 0 respectively contribute about 43 %.



**Figure 12: Top revenue generating job codes**

***Ticket Revenue***:

This column helps understand revenue generated from each individual ticket in conjunction with branch, customer type and job code.

***Call date, schedule date, dispatch date, and complete date:***

These columns help prepare analysis of trend, seasonality, yearly, monthly and day of week, etc. for each branch, customer type and job code.

***Call time, schedule time, dispatch time, and complete time:***

These columns help understand and derive Time of Day analysis, avg. service time for jobs across branches, also disgruntled customers.

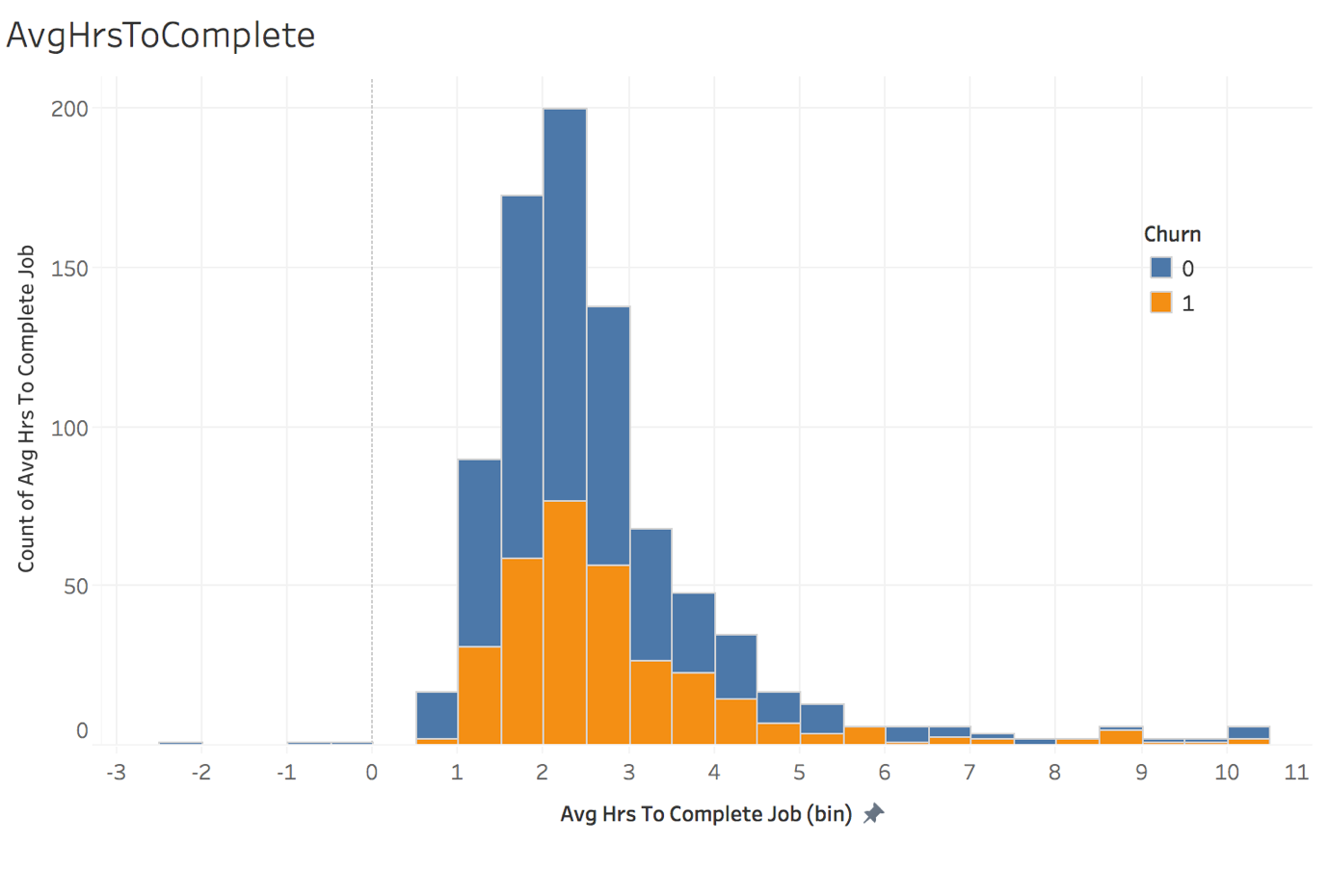
***State/City:***

Gives geographical information of customer and will contribute in branch/state level analysis

***Churn:***

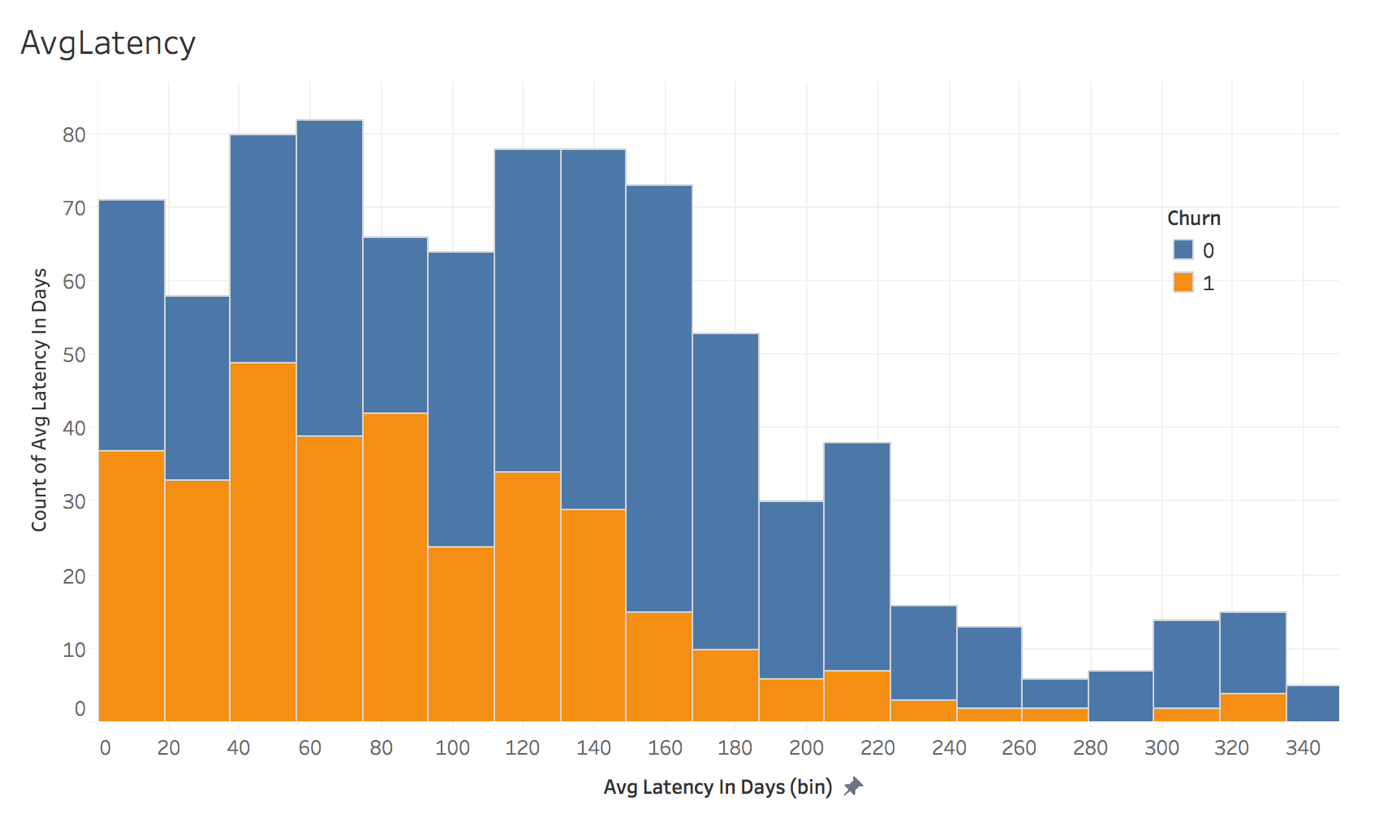
Customer who did not show in last 6months are flagged as lost/churned customers. Based on this we defined Churn variable for our training and test data.

***Univariate Analysis of some derived features against churn:***



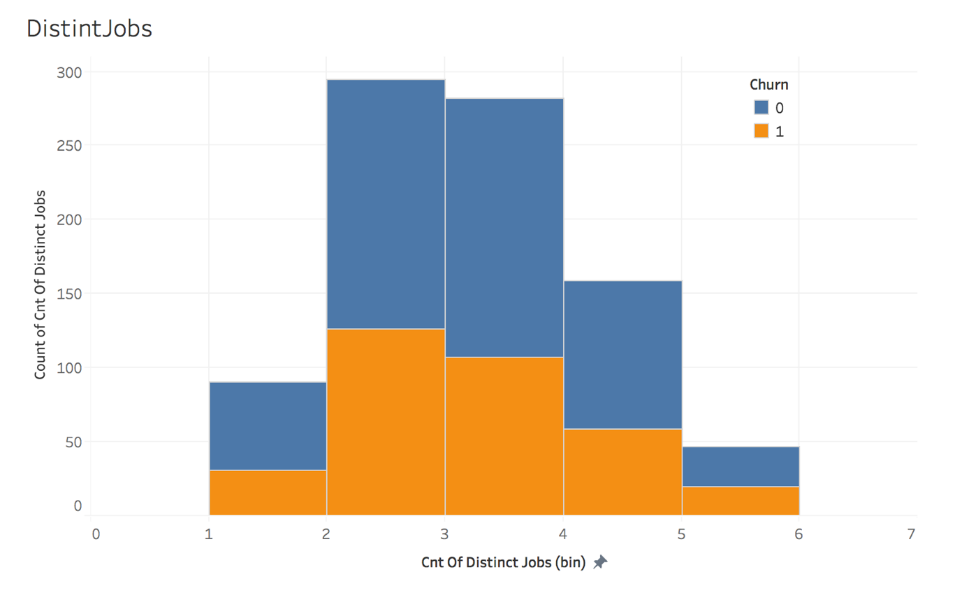
**Figure 13: Average time (hours) taken to complete the job**

To understand the customers that have churned, a univariate analysis of different variables that are expected to influence churn versus customers that have churned is carried out. The first variable studied is churn based on the average number of hours taken to complete the work. The data is shown in Figure 13. The highest number of calls (200) are received for calls that require between 2 and 2.5 hours to be completed of which about 75 calls (30%) of the customers churn. The highest percent churn (100%) however is for jobs that take between 5.5 to 6 hours to complete.



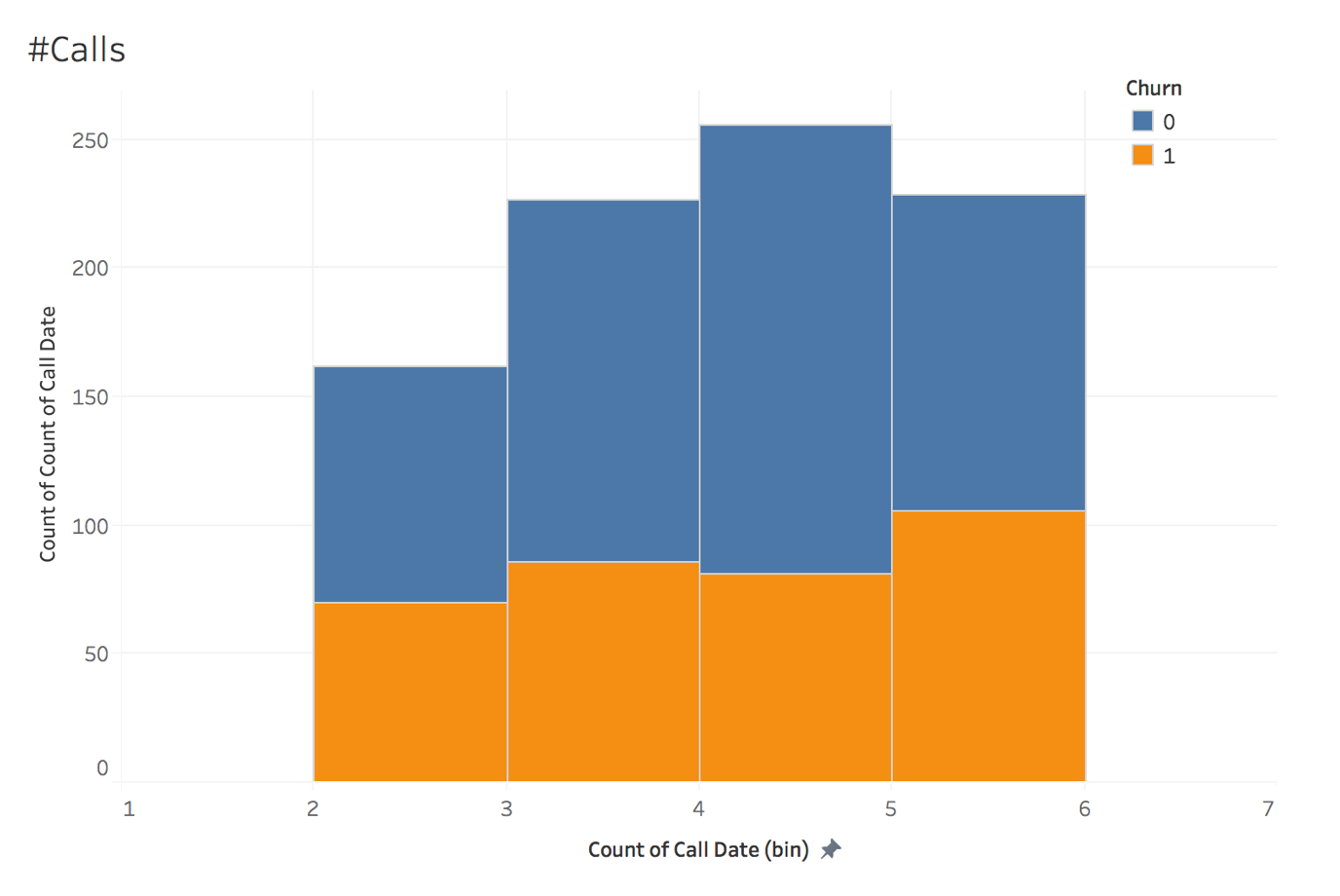
**Figure 14: Churn data for latency (days)**

Latency, in this case, is defined as the delay in (days) between consecutive service calls. Latency is expected to influence churn, and data of average latency in days versus churn is shown in Figure 14. In general, higher churn is visible for lower latency days. This can be attributed to domain or demand of plumbing service business. The highest churn occurs for a latency period between 40-80 days. The most loyal customers also belong to this area as well.



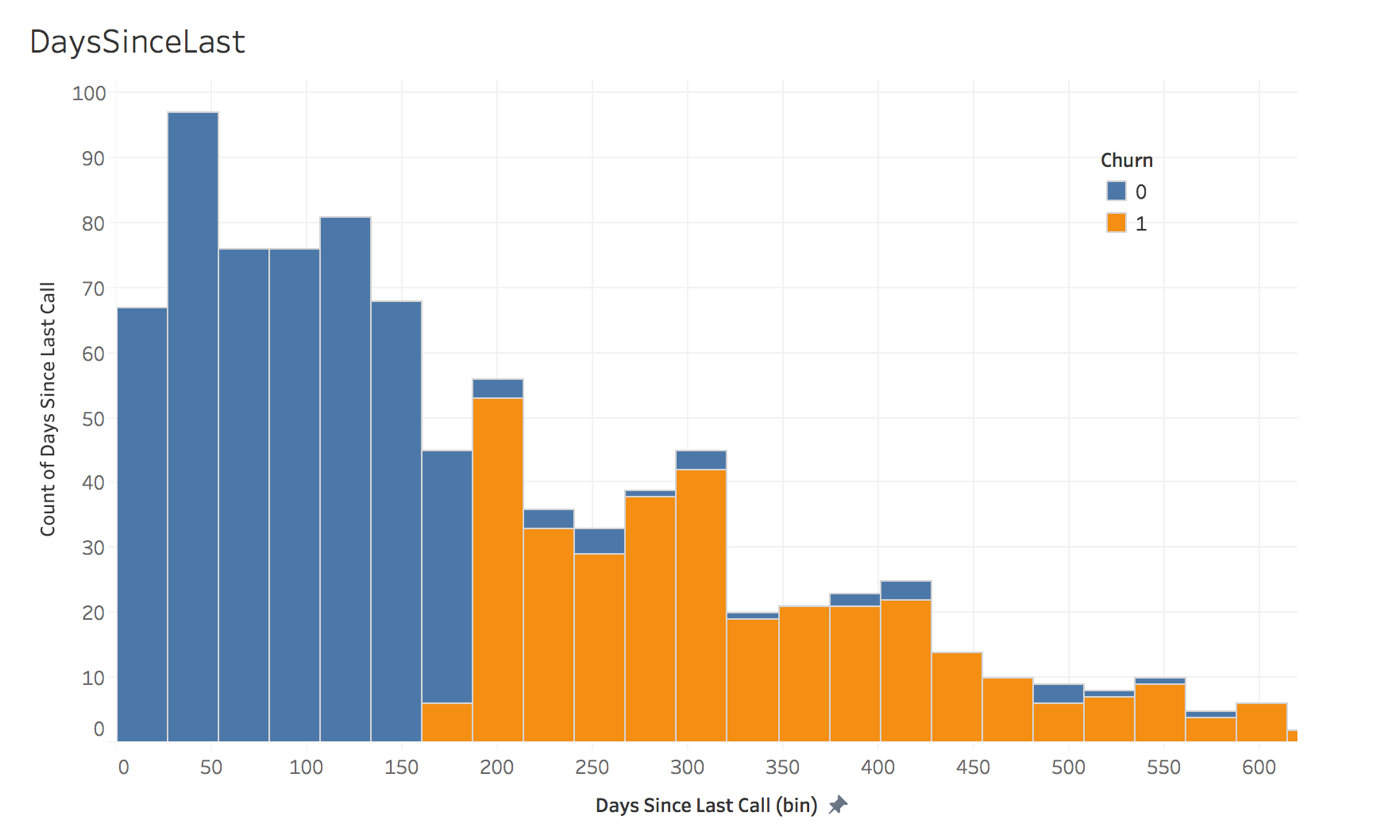
**Figure 15: Churn by customer’s count of distinct job codes**

Similarly, if we observe churn over customer’s count of distinct jobs. Customers used service for more than 1 different jobs happen to churn as shown in Figure 15. Churn is observed highest when count of distinct jobs is 2 and 3.



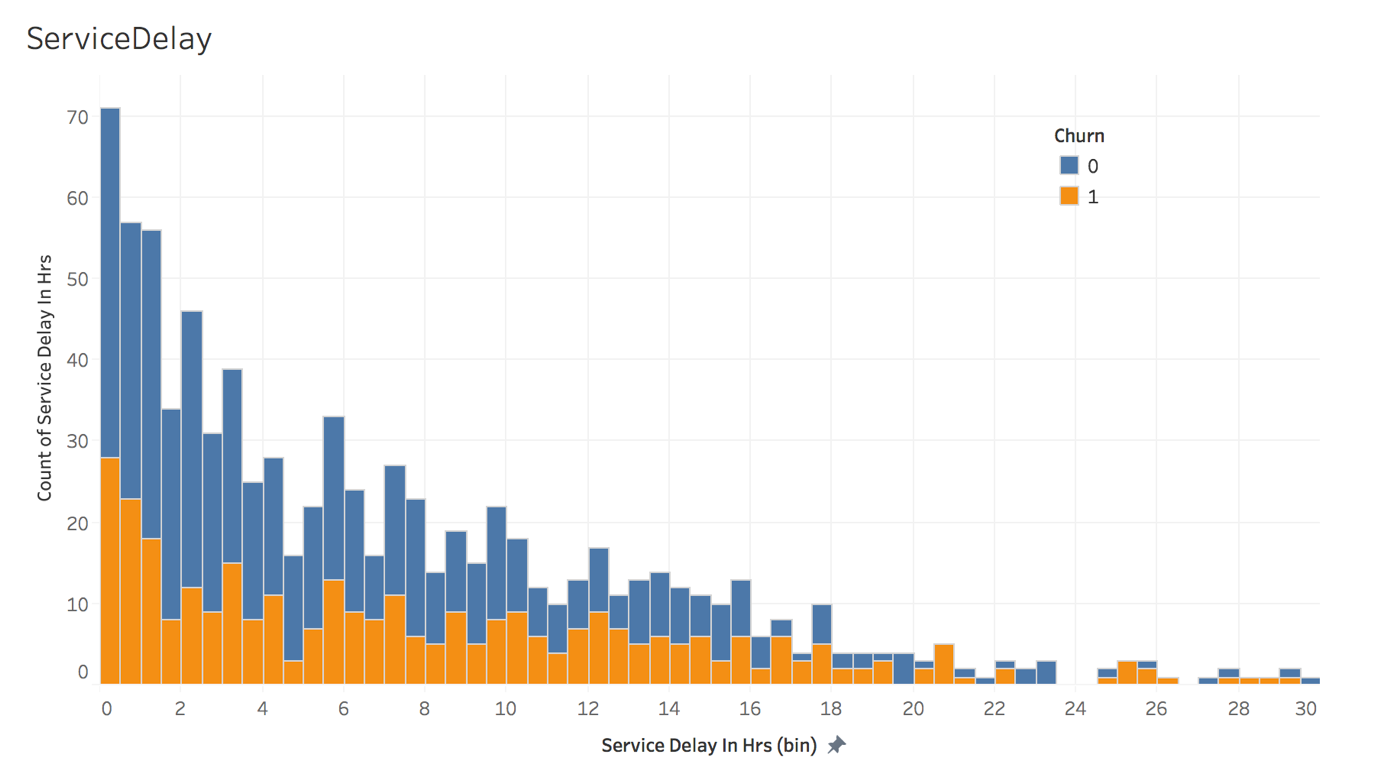
**Figure 16: Churn by number of calls from customer till date**

This is one of the most useful insights, Figure 16 shows churn over number of service calls made by customers till date. In general churn increases with increase in number of calls. This also can be attributed to nature of plumbing business and is in sync with Average Latency vs Churn plot. Churn is observed highest in customers who called maximum times. As count of service call by customer increases churn also increases.



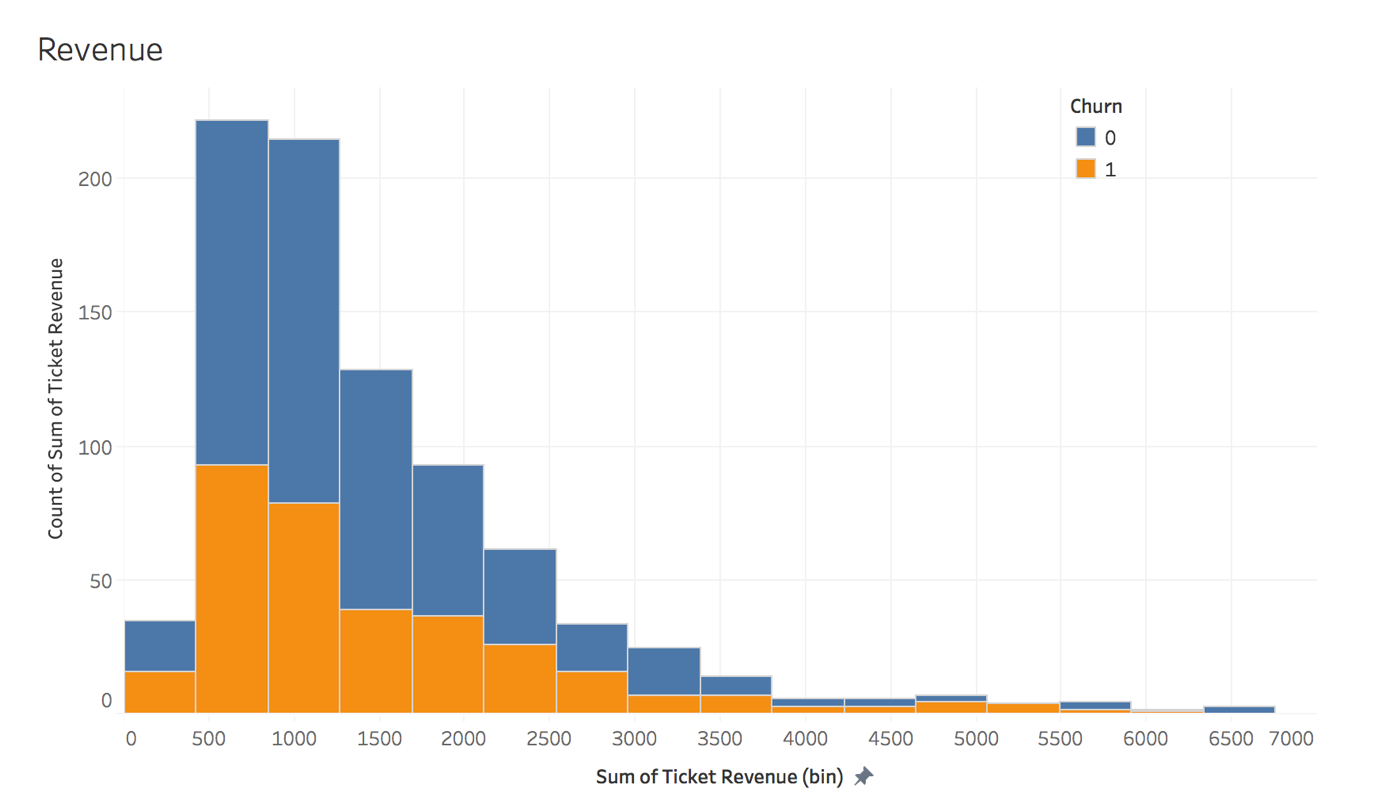
**Figure 17: Churn by number of days elapsed since last call date**

Figure 17, shown below, is one of the key feature defining our churn rule. Here, churn is highlighted against bins of number of days that have passed since the last service call by customer. Figure 17 shows churn behavior over days since last called for service. It follows that churn is significant for customers not called for more than 150days.



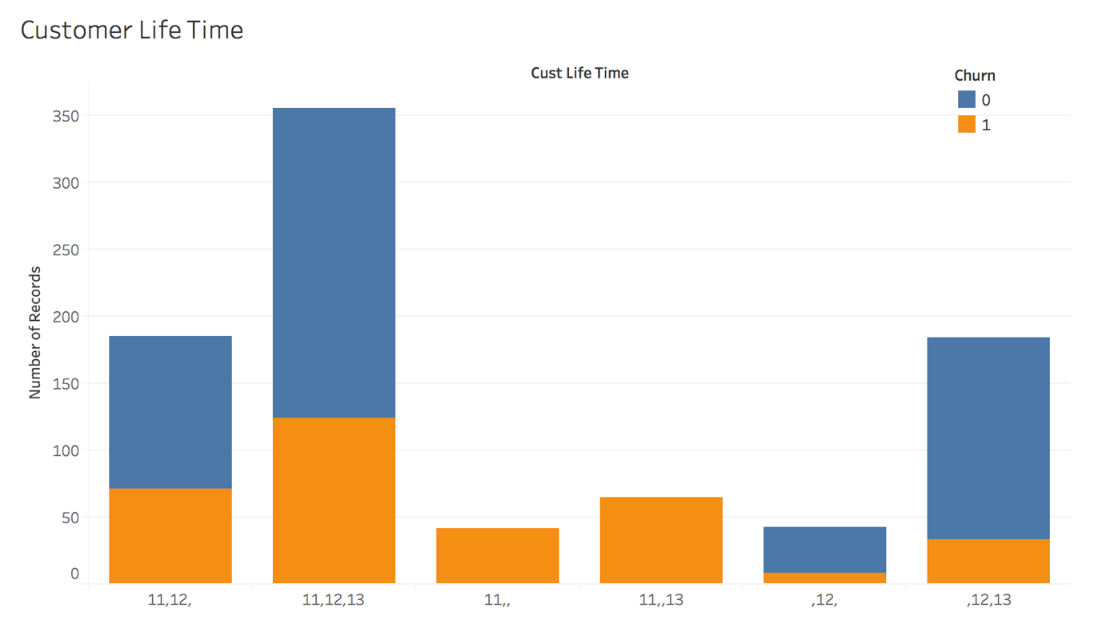
**Figure 18: Churn by sum of service delay in(hrs.)**

Churn data as a function of sum of service delay in hours throughout the journey of individual customer is given in Figure 18. It is generally expected that churn increases with increase in delay. However, contrary to expectations, it is shown that churn is highest for delay between 0 to 1 hours. It is speculated that the churn happens due to the urgent nature of the task. For example, if a customer requires an urgent task to be fixed, and if an associate of the company is not available during that time, then the customer looks for an alternate vendor to attend to the problem. Purely dependent on domain of plumbing business.



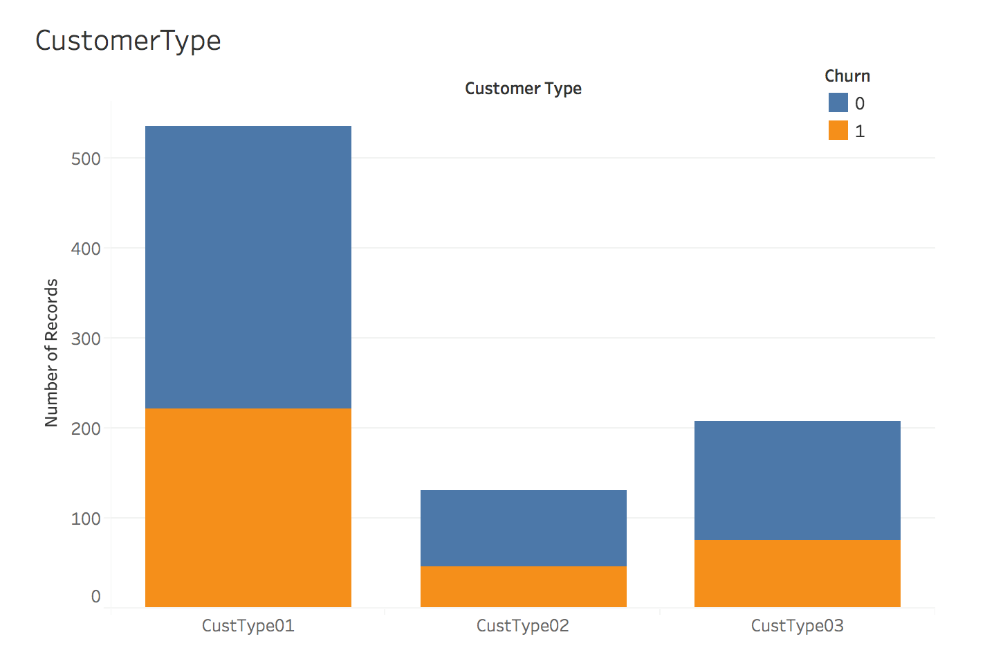
**Figure 19: Churn by sum all service ticket revenue for individual customer**

Here, churn is highlighted on top of sum of all service ticket’s revenue for individual customer. It can be inferred from Figure 19 that customers with lower sum of all ticket revenue tends to churn more. Highest churn occurs for sum of ticket revenues between $500 to $ 1000 and as sum of revenue increases churn decreases.



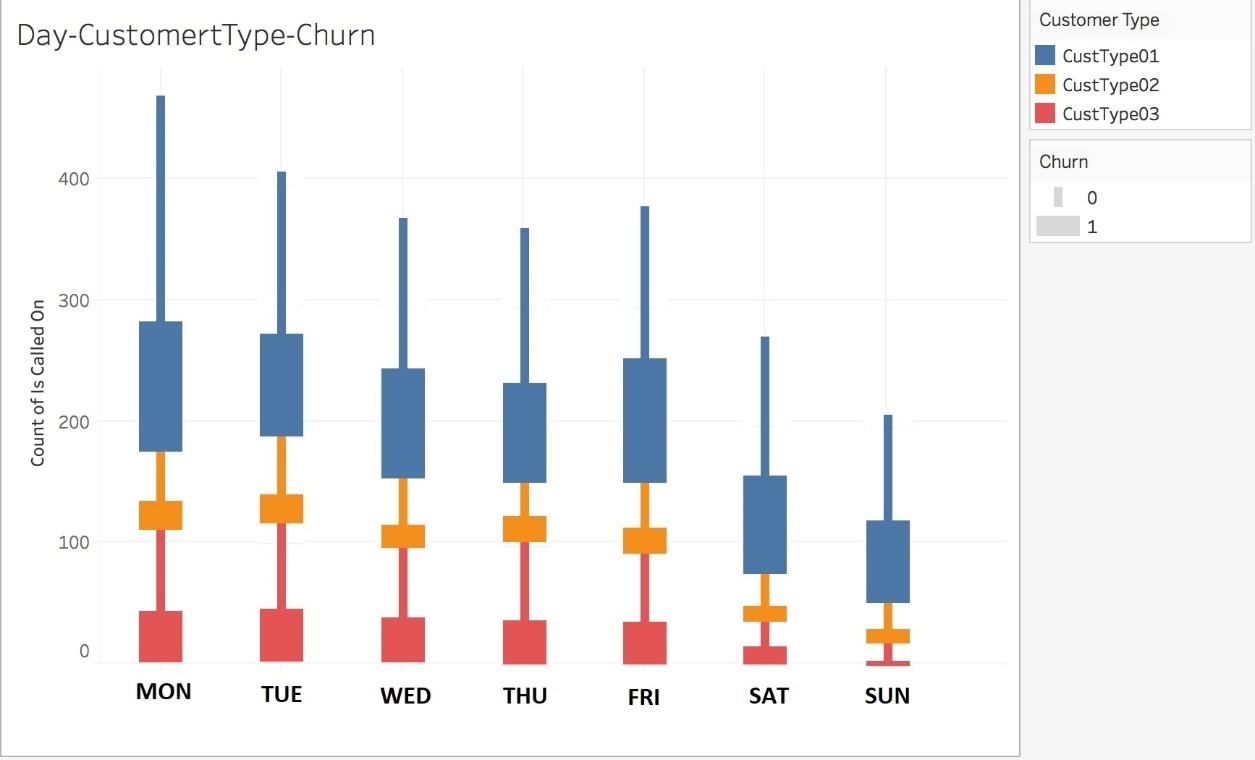
**Figure 20: Churn by customer lifetime**

Churn data as a function of customer life time is shown in Figure 20. Customer life time is nothing but a tracker which gives the count of all years that a given customer called for service. As evident from the figure, the higher the number of years the customers call in for services, lower is the probability of churn. In general, a lower percentage of churn occurs when customer call for service in consecutive years. Highest percentage (100%) of churn occurs for 2011 customers who do not return for services. We can also deduce from here that, every year new customers are acquired, and few customers are lost. However, the plumbing services company has a sizeable number of returning or loyal customers.



**Figure 21: Churn data by customer type**

Churn by customer type is shown in Figure 21. As expected the customer type 1 makes the highest number of calls and has the highest churn among all customer types.



**Figure 22: Churn data by customer type on day basis**

The correlation between churn and customer type as a function of day of call in depicted in Figure 22. Monday happens to be most busy day of business, also the customers who call on Monday happen to churn more.

**Figure 23: Percentage of yearly new versus repeat customers**

Figure23 shows percentage yearly average of new versus returning customers. 63% out of total calls are made by existing/returning customers whereas 37% of calls are made by new customers. Out of 37%, majority of customers have been converted to repeat customers.

## 2.3 Data Dictionary

Explanations for the variables are presented in Table 6. Each of the variables is assigned a best (primary) category.

**Table 6: Explanation of variables present in the data set**

|  |  |  |  |
| --- | --- | --- | --- |
| Placement (Col no) | Data Header | Type | Definition |
| 1 | **Branch ID** | String | ID of the branch where the request for service was received |
| 2 | **Customer ID** | String | ID allocated to the customer to track customer behaviour |
| 3 | **Cust Life Time** | Numeric | Years in which the customer had placed for a service request i.e. 11/12/13 |
| 4 | **Bill To** | Numeric | To whom the bill must be raised on |
| 5 | **Customer Type** | String | Given customer Id belongs to which type: Type1, type2 or type3 |
| 6 | **Customer Name** | String | Name of the customer |
| 7 | **Address** | String | Address of customer |
| 8 | **City** | String | City where customer is located |
| 9 | **State** | String | State where customer is located |
| 10 | **Zip Code** | String | Postal Zip code of the customer |
| 11 | **Zip 5** | String | First 5 digits of the Zip code |
| 12 | **Area 1** | String | County in which the customer is located |
| 13 | **Area 2** | String | Town/ District in which the customer is located |
| 14 | **Contact** | String | Contact person’s name of the customer |
| 15 | **Setup Date** | Date | Date when the customer ID was setup |
| 16 | **Last Service Date** | Date | Date when the customer last availed of service |
| 17 | **Ticket Number** | Numeric | Ticket number allocated to the service request |
| 18 | **Current Email** | String | Email address of the customer currently registered |
| 19 | **Rev Code** | String | Revenue code |
| 20 | **Job Code** | String | Job code of the service request from customer |
| 21 | **Ticket Revenue** | Numeric | Revenue generated from the ticket |
| 22 | **Year Month** | Numeric | Year and month of the ticker denoted in YYYYMM |
| 23 | **Year** | Numeric | Year of service |
| 24 | **Week Ending Date** | Date | Date of the day when the week ends |
| 25 | **Call Date** | Date | Date when the service ticket was raised |
| 26 | **Call Time(Master)** | Character | Time when the ticket is raised in XXXX hrs |
| 27 | **Call Time** | Character | Time when the ticket is raised in HH:MM format |
| 28 | **Rounded Call Time** | Character | Rounded off time for Call Time to the next hour in HH:MM format |
| 29 | **Schedule Date** | Date | Date when the job is scheduled to be completed for the ticket |
| 30 | **Schedule Time (Master)** | Character | Time when the ticket is scheduled to be completed in XXXX hrs |
| 31 | **Schedule Time** | Character | Time when the ticket is scheduled to be completed in HH:MM format |
| 32 | **Schedule Time** | Character | Time when the ticket is scheduled to be completed in HH:MM format in AM/PM |
| 33 | **Rounded Schedule Time** | Character | Rounded off time for Call Time to the next hour in HH:MM format |
| 34 | **Dispatch Date** | Date | Date when the team from the service provider was dispatched for the job |
| 35 | **Dispatch Time(Master)** | Character | Time when the dispatch team carried out the job in XXXX hrs format |
| 36 | **Dispatch Time** | Character | Time when the dispatch team carried out the job in HH:MM format |
| 37 | **Rounded Dispatch Time** | Character | Rounded off time for the dispatch team to the next hour in HH:MM format |
| 38 | **Complete Date** | Date | Date when the ticket / job is completed |
| 39 | **Complete Time(Master)** | Character | Time when the job is completed in XXXX hrs format |
| 40 | **Complete Time** | Character | Time when the job is completed in HH:MM format |
| 41 | **Rounded Complete Time** | Character | Rounded off time for the Completed time to the next hour in HH:MM format |
| 42 | **Days to Complete** | Numeric | No. of days taken to complete the job |
| 43 | **Days Prior schedule** | Numeric | No. of days the job was completed prior to the scheduled completion date |

## 2.4 Data Quality Concerns and Preparation

There are couple of concerns majorly with quality and quantity of data. Since data is live and being provided by actual company for academic purpose we could not negotiate much. We could have performed even better analysis if were provided data from different channels of business and more quantity of transactions. Apart from quantity, there were around 300+ records with incorrect date details and being corrected as a part of data preparation phase.

## 2.5 Data Preparation

***Data Summary:***

1. We have used Plumbing Master Data V0.1.xls file for our analysis
2. This file has got total 31 variables of mixed datatypes
3. Out of 31 variables, 7 categorical, 7 Date, 4 time variables, and others masked customer
4. information variables
5. Total 5284 records/observation
6. Observations were spread from year 2010 to 2014

***Variables Transformation:***

1. Cust Life Time: Derived from Call date years given customer used service
2. Count of Distinct Jobs: Number of different service Jobs used by customer
3. Days Since Last Call: Derived from finding number difference between (Today's date: last service date)
4. Service Delay in Hrs.: Derived by finding Sum of total service delay in hours for every customer (Dispatch time minus Schedule time)
5. Avg. Hrs. To Complete Job: Derived by finding average hours taken to complete job for given customer
6. Avg. Latency in Days: Derived by finding average number of days between calls for given customer
7. Is Called On DAY: Dummy binary variable to flag customer called on a day. All 7 days of week.
8. Sum of Ticket Revenue: Driven by finding Sum of total revenue from all tickets from given customer
9. Count of Call Date: Number of service calls made by given customer
10. Churn: Derived based on churn rule whether customer has churned or not: 1 – churned, 0 - Returning

***Missing values and Outliers:***

1. Missing/misleading value treatment: Treated/replaced records~500 having Schedule time as 5AM with Avg. difference between schedule call time and Dispatched calls time
2. Also, treated Call date/time, Schedule date/time, dispatch date/time and complete date/time and last service date as per their order for incorrect records.
3. Outlier treatment: Fixed records having incorrect date/year based on rest date and time features. e.g. Call date 10/02/2011 and complete date 11/04/2019 etc.
4. Converted all date and time variables in correct form and added Rounded time dummy variable for Day of week and time of day analysis.
5. Dropped 212 records of year 2010 and 2014.

***Data partition:***

Creating Development and Holdout Sample

1. Development(Train): All call records in year 2011 and 2012
2. Holdout(Test): All call records in year 2013

# 3.0 Hypothesis Testing

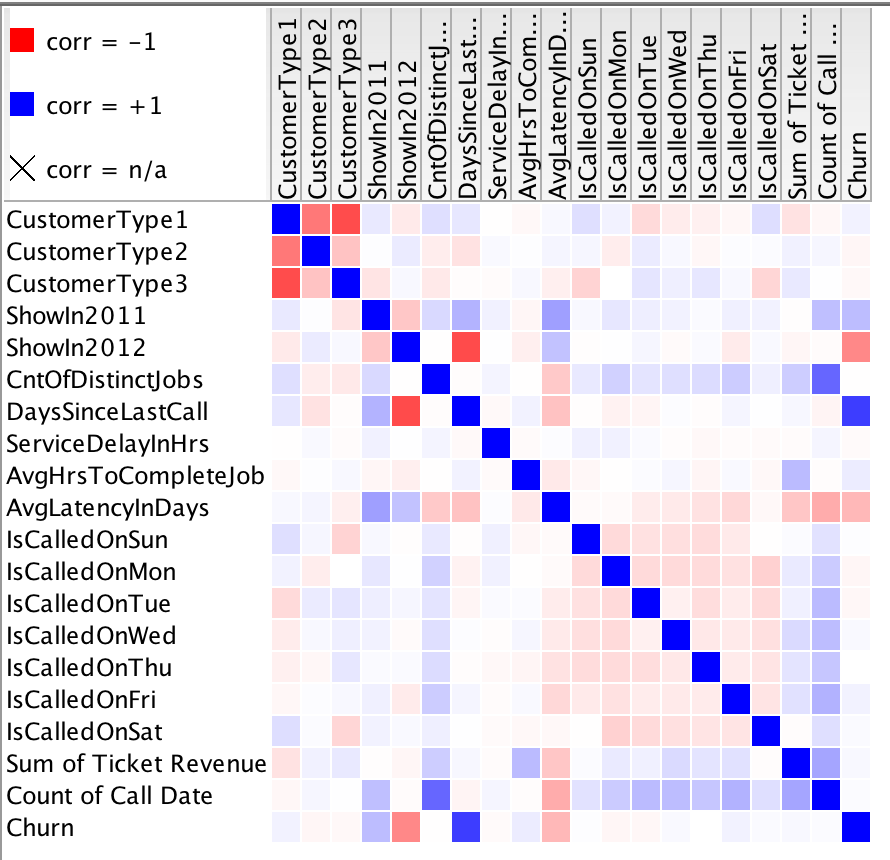
## 3.1 List of Null hypotheses:

**Table 7: List of Null Hypothesis statements**

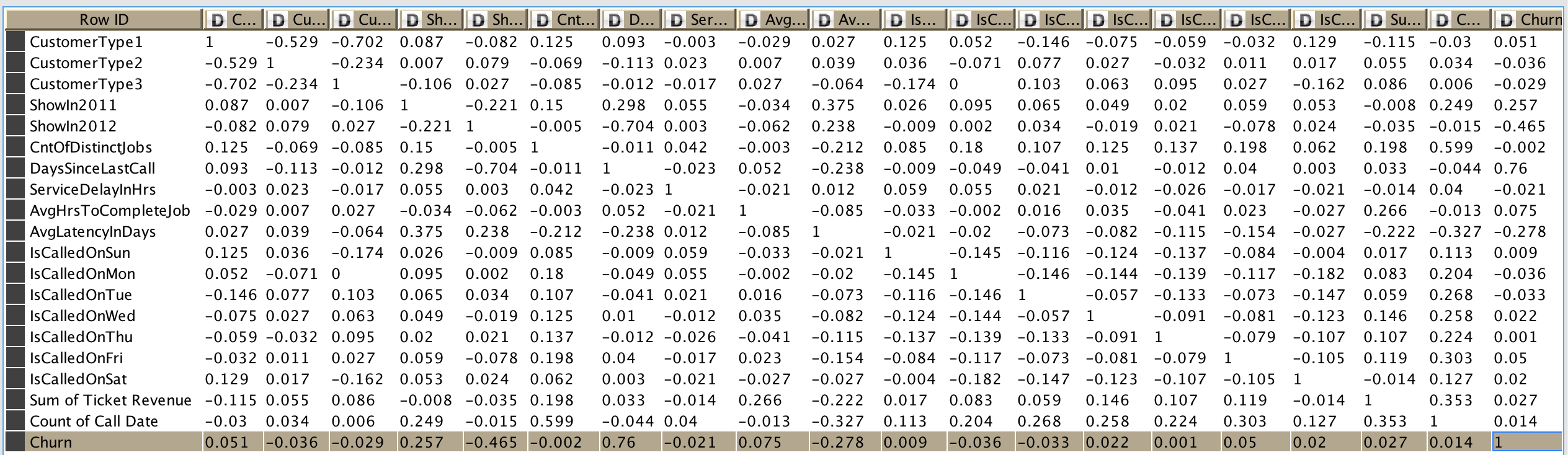
|  |  |  |
| --- | --- | --- |
| No. | Null Hypothesis statements | \*Status |
| 1 | None of the derived features contribute to predict customer churn. | Rejected |
| 2 | Average number of hours to complete job increases probability of customer churn increases | Accepted |
| 3 | Average Delay in days between service calls i.e. Avg. Latency has direct influence on customer churn | Accepted |
| 4 | More the number of distinct jobs used by customer, lesser the probability of customer churn | Rejected |
| 5 | More the number of service calls made by customer, lesser the probability of customer churn | Rejected |
| 6 | As days since last call increases, probability of customer churn increases | Accepted |
| 7 | More the delay to pick up scheduled service calls, higher the probability of customer churn | Rejected |
| 8 | Higher the total revenue from customer, lower the probability of customer churn | Rejected |
| 9 | service call day of a week does influence customer churn | Rejected |
| 10 | If customer uses serviced in consecutive years, then there is less chance that the customer churns | Accept |

\*Hypothesis Status is updated based on Hypothesis tests performed as shown below (correlation matrix and linear regression) and exploratory analysis carried out as part of Chapter 2.

## 3.2 Correlation matrix

****

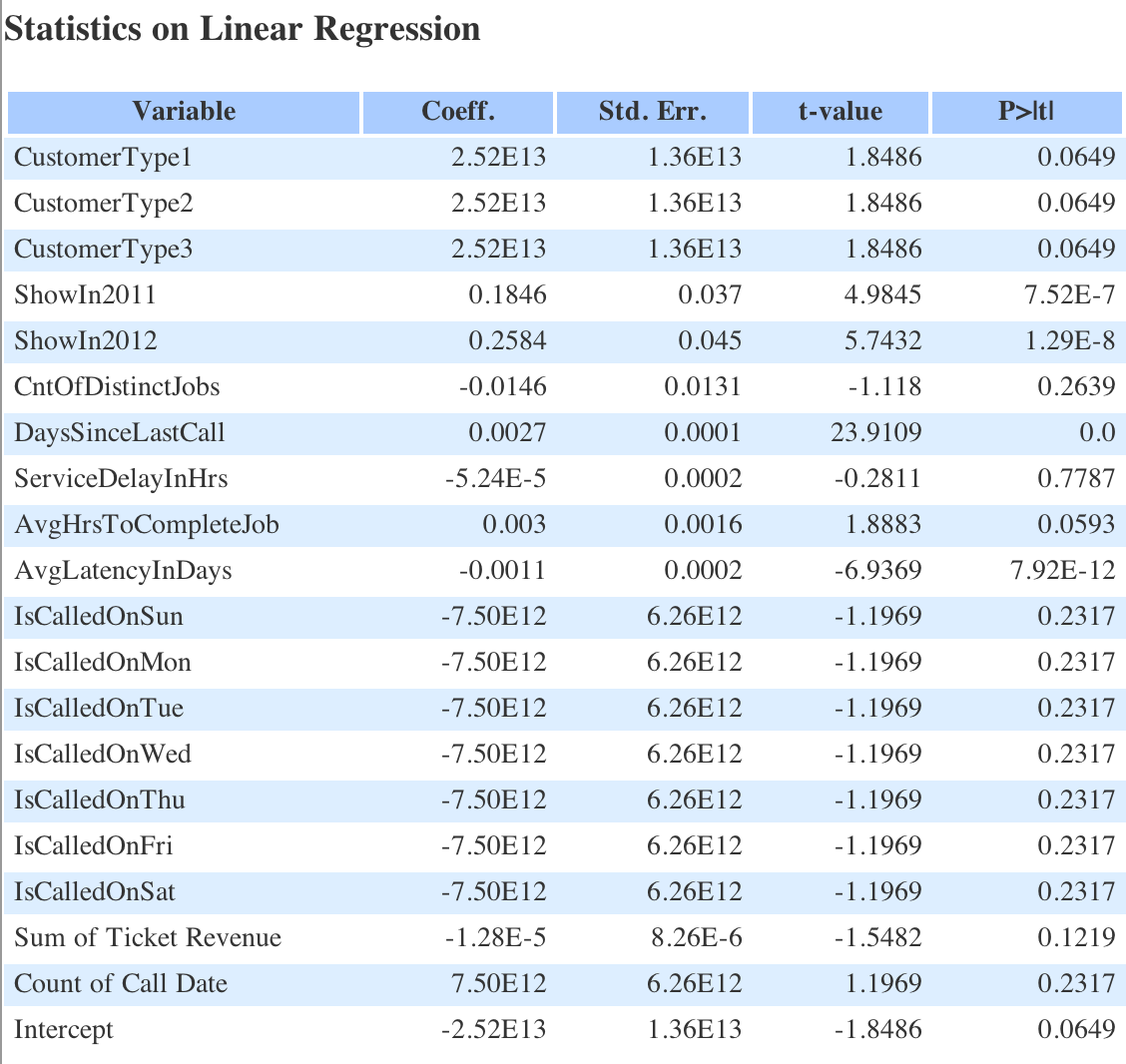
**Figure 24: Correlation matrix**

****

**Figure 25: Output for correlation matrix**

## 3.3 Hypothesis testing using linear regression equation

**Table 8: Statistics on Linear Regression**



## 3.4 List of Significant Variables for Model Building

Based on Correlation matrix and Linear regression statistics the variables found to be significant for POC and model building is shown in Table 9.

**Table 9: Significant variables for model building**

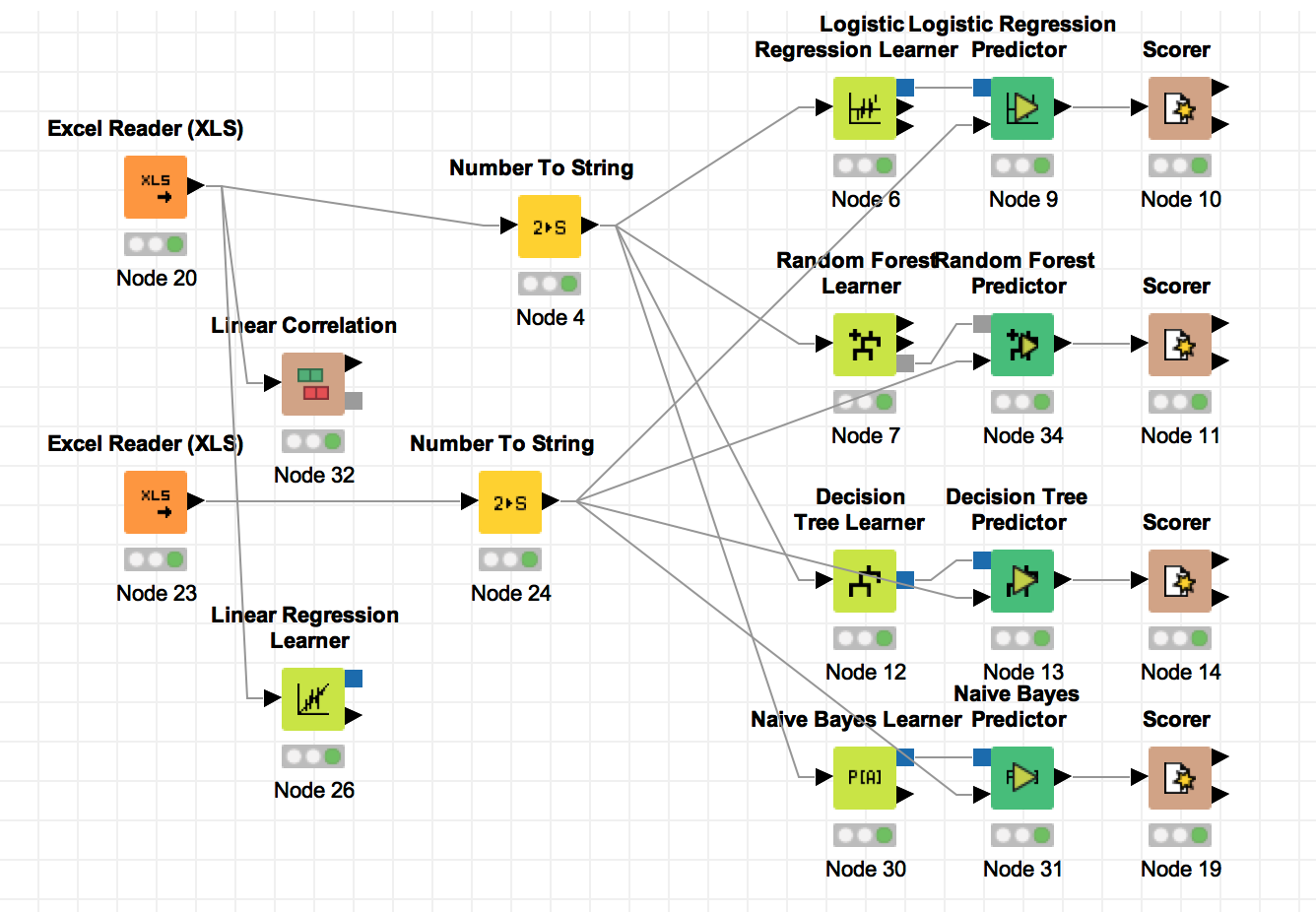
|  |  |
| --- | --- |
| Variable | p-Value |
| ShowInPrevYear | 7.25E-07 |
| ShowInCurrYear | 1.29E-08 |
| DaysSinceLastCall | 0 |
| AvgHrsToCompleteJob | 0.05 |
| AvgLatencyInDays | 7.92E-12 |

# 4.0 Modelling

## 4.1 Proof of concept(POC)

We have used KNIME for our POC and built 4 different models on cleaned training data and evaluated their prediction performance. We used the 5 variables that were shown to be significant during our hypothesis testing studies and tested them using the following 4 models to predict churn.

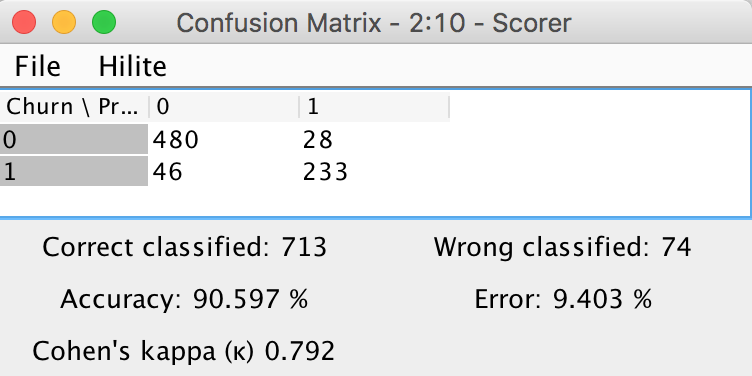
1. Logistic Regression
2. Random Forest
3. Decision Tree
4. Naïve Bayes



**Figure 26: Knime output showing model building approach**

## 4.2 Logistic Regression

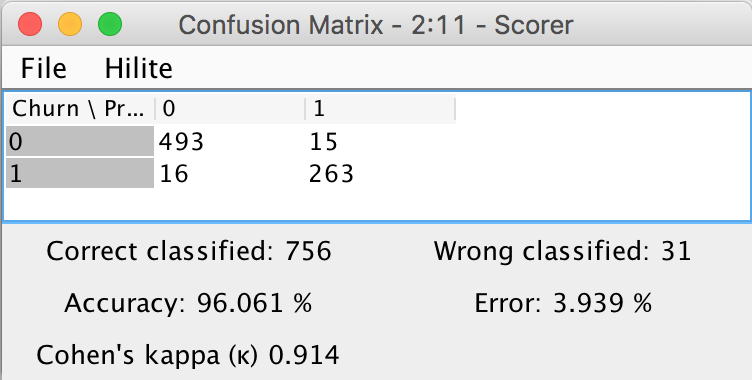
The confusion matrix from the Logistic Regression model is shown in Figure 27. The accuracy of the model is 90.6 % and 74 observations are misclassified.



**Figure 27: Confusion matrix of the Logistic Regression Model**

## 4.3 Random Forest

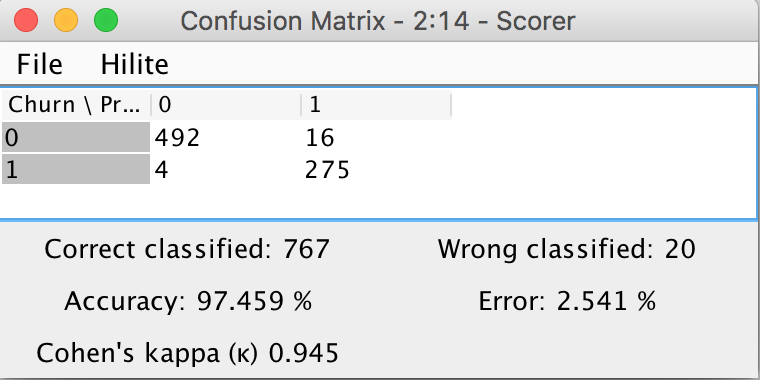
The confusion matrix from the Logistic Regression model is shown in Figure 28. The accuracy of the model is 96 % and 31 observations are misclassified.



**Figure 28: Confusion matrix of the Random Forest Model**

## 4.4 Decision Tree

The confusion matrix from the Logistic Regression model is shown in Figure 29. The accuracy of the model is 97.5 % and 20 observations are misclassified.



**Figure 29: Confusion matrix of the Random Forest Model**

## 4.5 Naive Bayes

The confusion matrix from the Logistic Regression model is shown in Figure 30. The accuracy of the model is 83.6 % and 129 observations are misclassified.



**Figure 30: Confusion matrix of the Naïve Bayes Model**

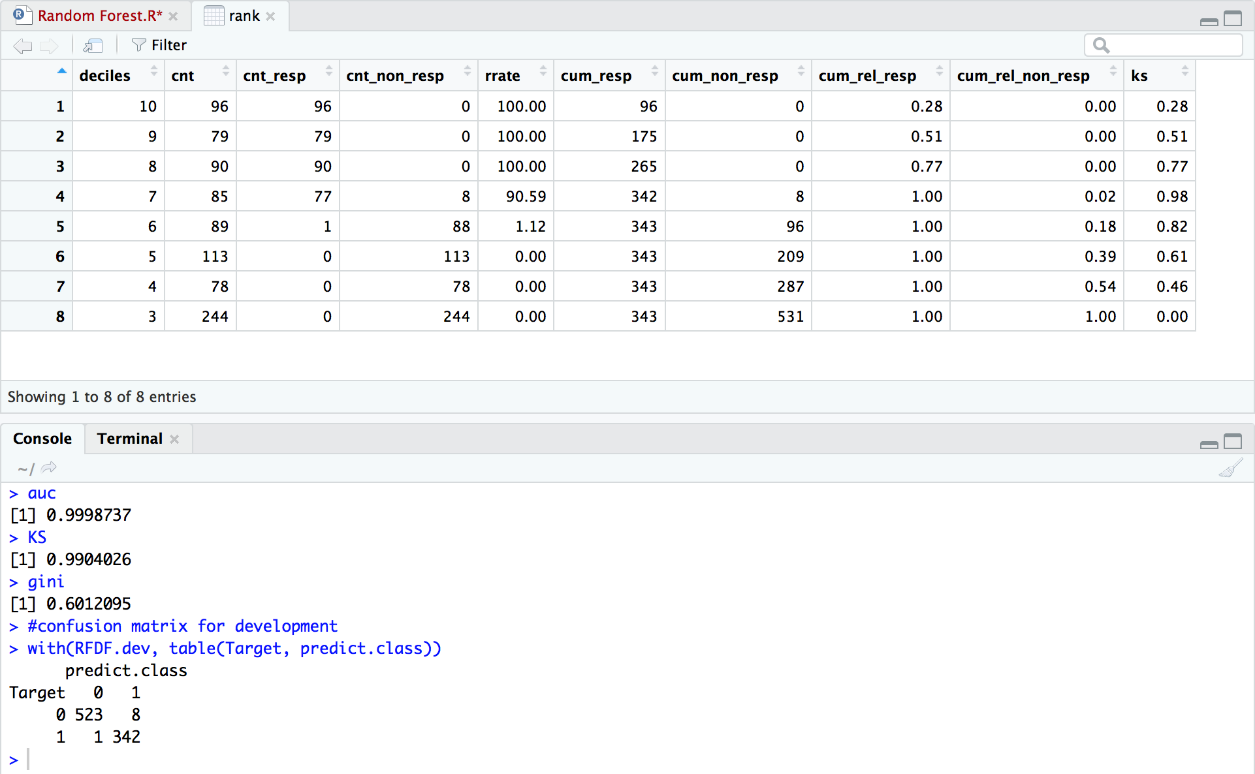
## 4.6 Final model using R Studio

Random forest was used to build the final model in R Studio. The following steps were followed while building the model to infer and then predict customer churn.

1. Full RF model was built using Train\_Model2 dataset prepared in Modeling\_Cleaned
2. The OOB chart is then plotted, followed by building of a pruned RF model using OOB
3. Deciles are created based on the predicted output followed by rank ordering
4. Calculated and printed Rank table, AUC, KS, Gini and confusion matrix on Train and then test data sets.

***Analyzing RF development model performance:***

Rank ordering, KS value, Confusion matrix of the full RF model are shown in Figure 31.

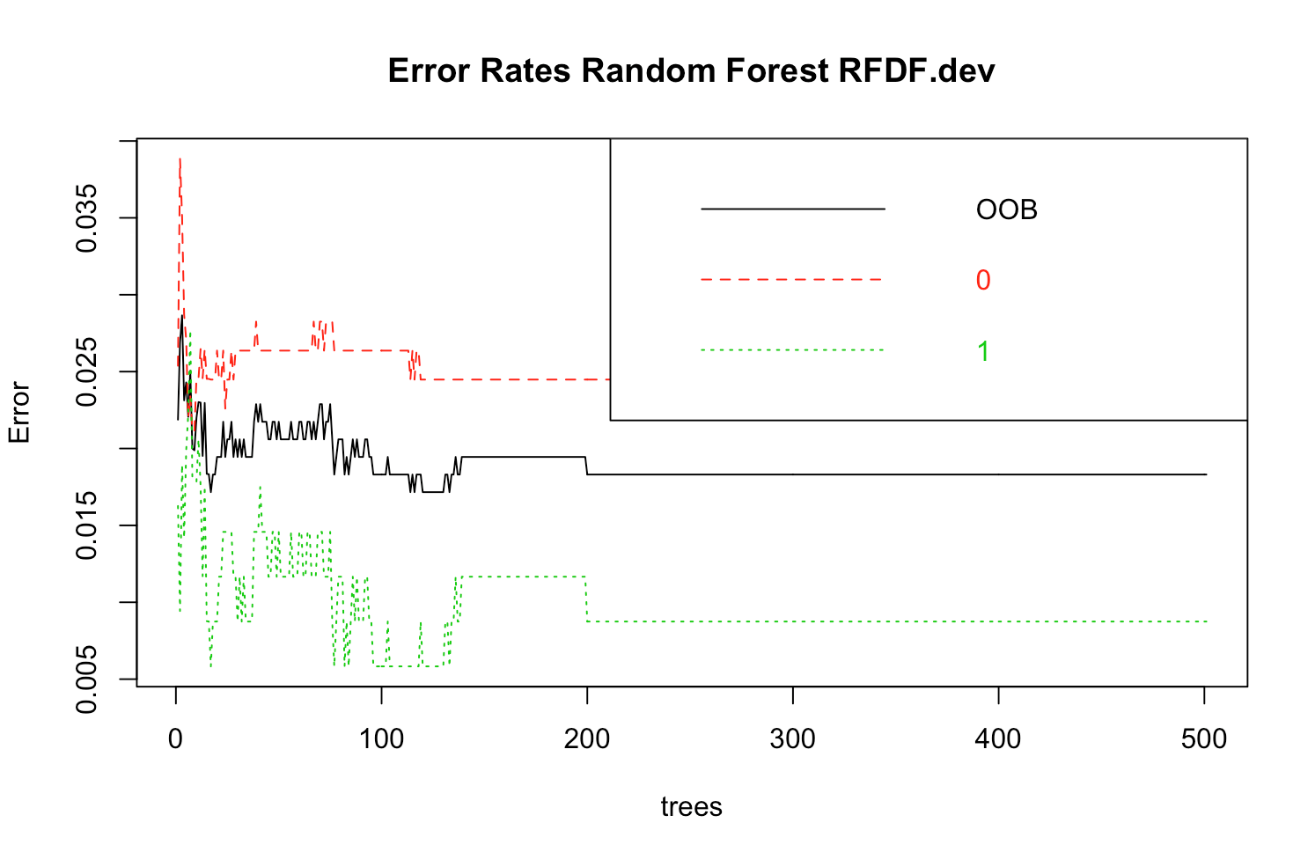


**Figure 31: Performance output of final RF model (development model)**

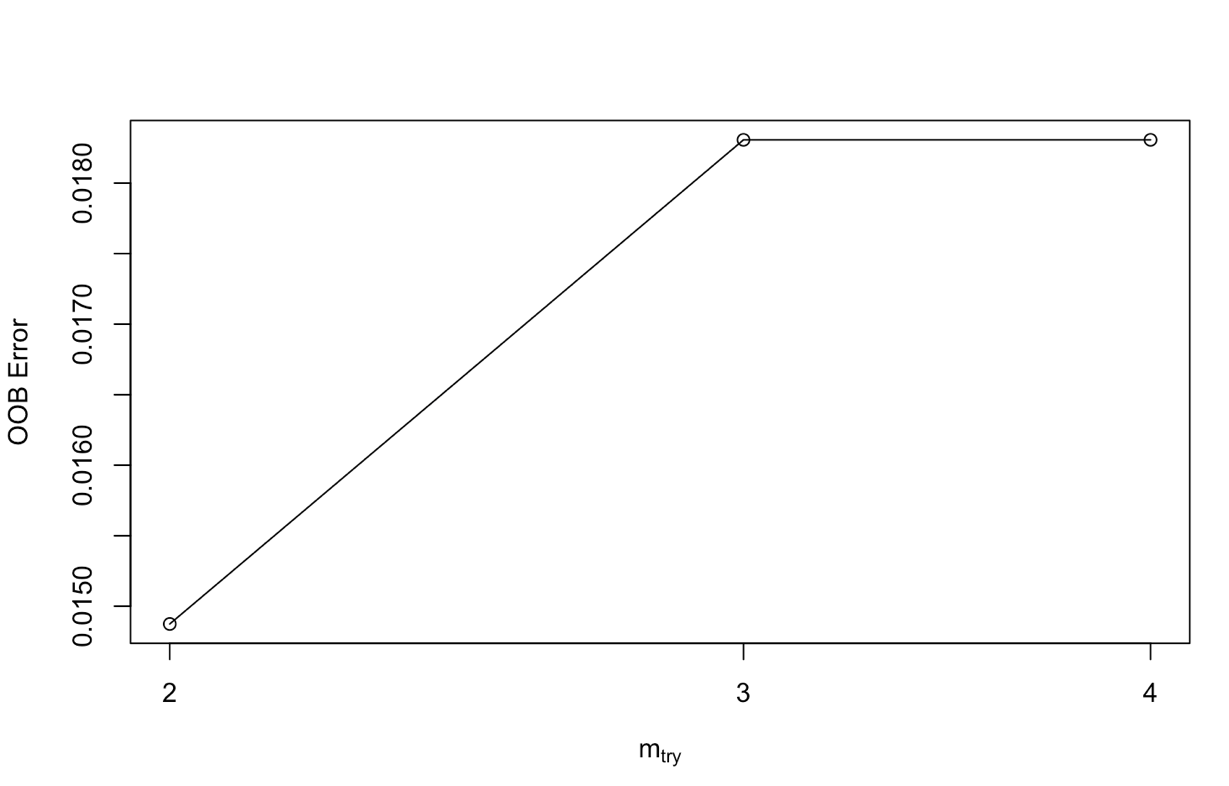
The performance measure of development model (rf1) is given below:

1. ***Rank table:*** Top 3 decile has 100% response rate i.e. 2.6times 39% overall response before applying RF algorithm.
2. ***Confusion Matrix:*** Correctness of prediction is 98.9%

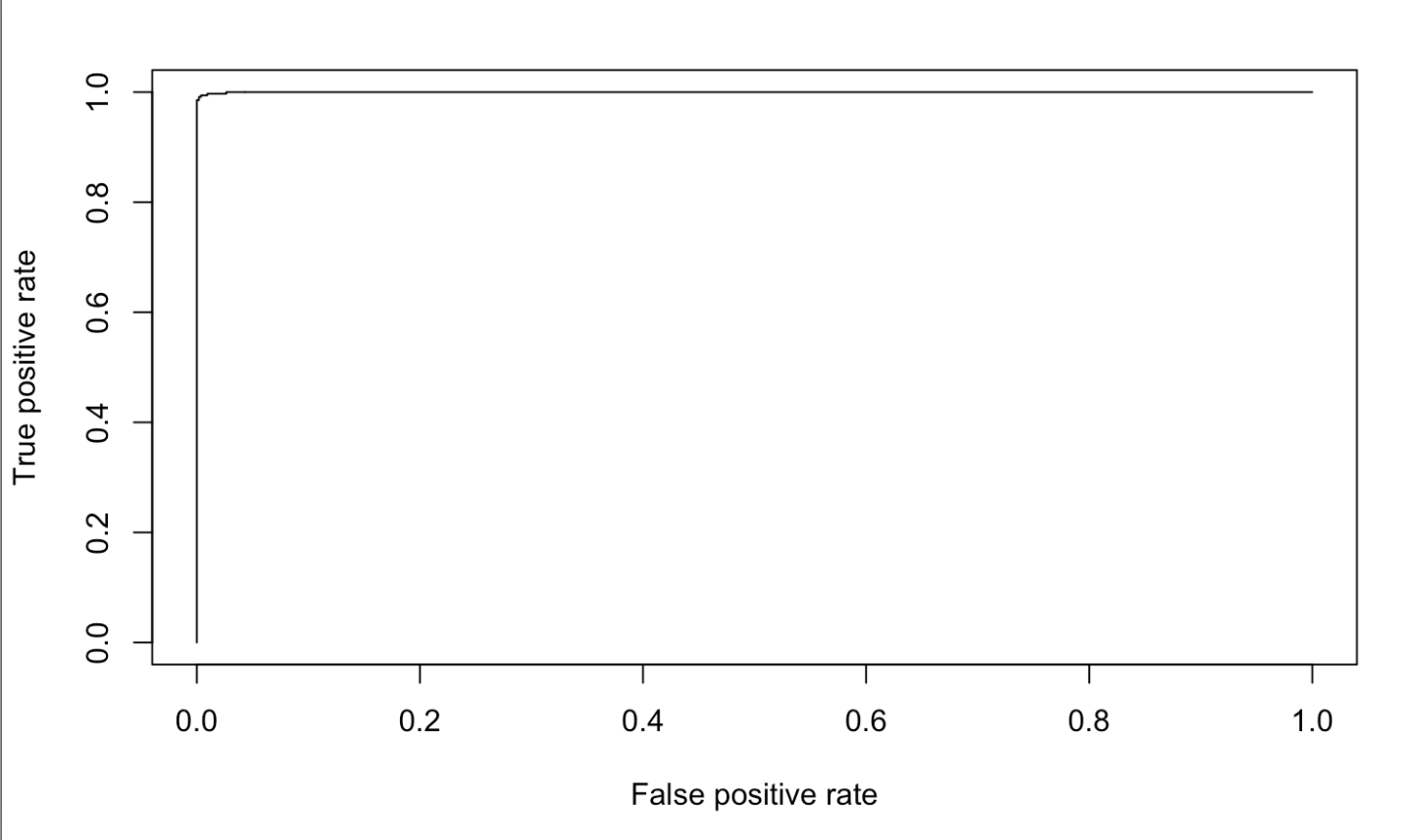
The OOB error plot, finding optimum using mTry curve, and AUC curve for the development model is shown by Figures 32, 33, and 34 respectively.



**Figure 32: OOB error rates of Random Forest development model**

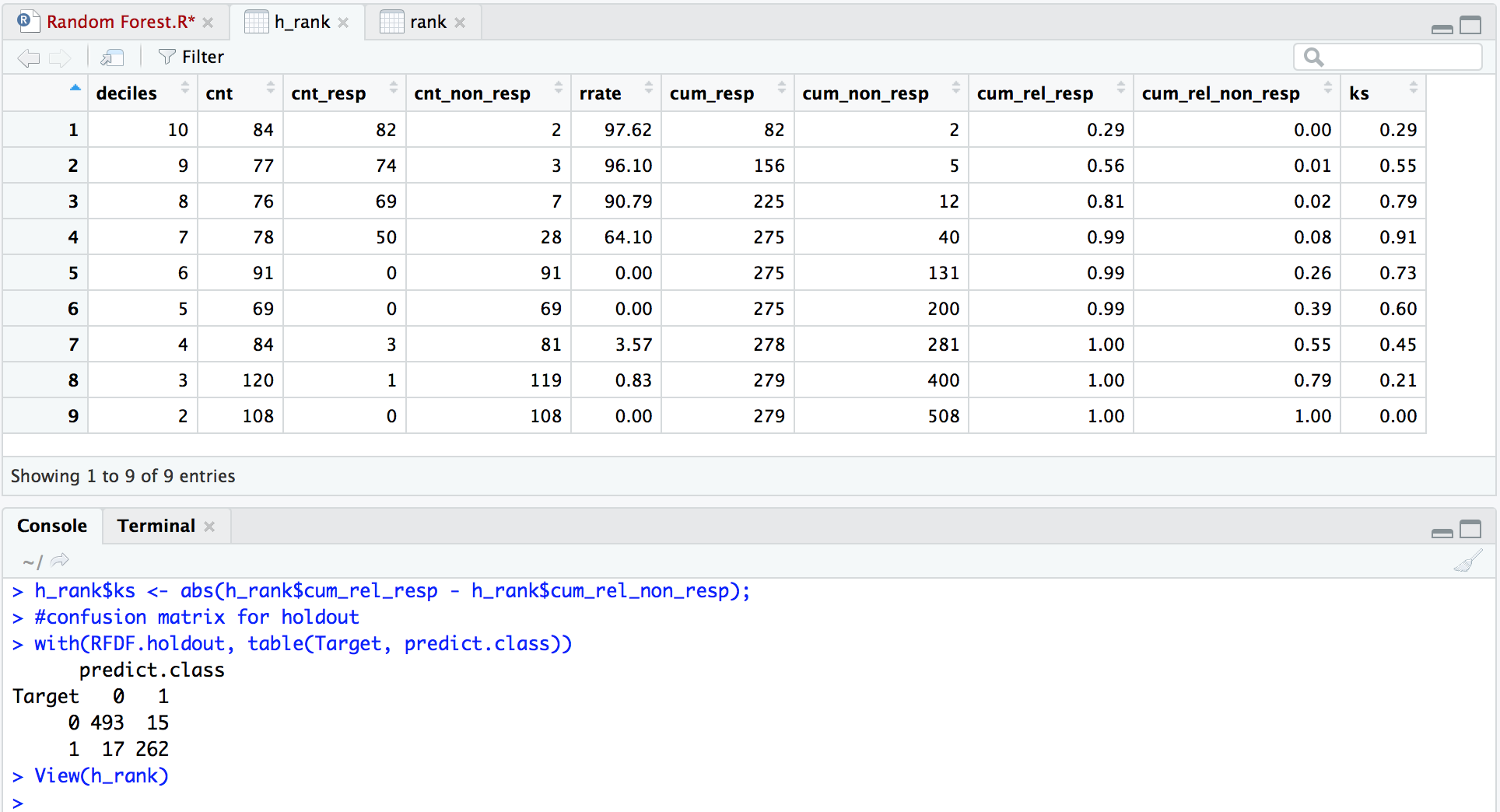


**Figure 33: mTry curve for finding optimum**

****

**Figure 34: AUC Curve of Random Forest development mode**

## 4.7 Analysing RF model on unseen holdout data

****

**Figure 35: Performance output of final RF model (holdout data)**

Rank ordering, Confusion matrix of the full RF model are shown in Figure 35. The performance measure of Random Forest Model on holdout data is given below:

1. ***Rank table:*** Top decile has 97.62% response rate i.e. 2.8 times 35% overall response before applying RF algorithm.
2. ***Confusion Matrix:*** Correctness of prediction is 95.9%.

# 5.0 Conclusions and Recommendations

## 5.1 Conclusions

1. Business is mostly flat throughout 2011 to 2013 with slight increasing trend
2. There are total 47 branches and avg. weekly calls is 33. We can say there are few branches who do not get even one single call in a week
3. There are few Jobs asked by customers most and they contribute major in total revenue.
4. Strange but real: This is the business where cream customers tend to churn mostly attributed to ad-hoc work requirements.

## 5.2 Recommendations

1. To retain cream customer: We recommend company to start Annual Maintenance contract program for their loyal customers. This is another way that the churn can be controlled to some extent
2. There are many Jobs hardly used and can be delisted from the services offered by the company
3. 63% of total service calls are from returning customers. We recommend up selling and cross selling to existing customers which will help increase the company’s share in customer wallet
4. There are several branches which contribute very less to the overall revenue. These branches can either be combined with other branches, try applying a few recommendations listed above to see if business improves, or closed the branch if it continues with its poor performance
5. The company also needs to build a better CRM system, find CLV values, prepare website with geographical details, accept online service requests, use chat bots and other services to improve performance

# Appendix

The complete R Code of the Random Forest Model is embedded in the document



# References

1. ) [https://www.ibisworld.com/industry-trends/market-research reports/construction/special-trade-contractors/plumbers.html](https://www.ibisworld.com/industry-trends/market-research%20reports/construction/special-trade-contractors/plumbers.html) (Accessed on August 19, 2017) [↑](#endnote-ref-1)
2. ) <http://nextjuggernaut.com/plumbers-on-demand.php> [↑](#endnote-ref-2)
3. ) <https://www.hyperlinkinfosystem.com/plumbers-on-demand-app-development-uber-for-plumbers.htm> [↑](#endnote-ref-3)
4. ) <https://www.urbanclap.com/delhi-ncr-plumbers?gclid=CjwKCAiAm7LSBRBBEiwAvL1-L6aIbPyFXhMn6ETG4Qqw7sp7m8Dq1HfFM9vM5CWPa9m60CZuuDBVGRoCmTIQAvD_BwE&utm_source=google&utm_medium=cpc&utm_campaign=Google-NCR-Cust-HomeServices-EL-Web-Search--service_booking_city_AG&utm_source=google&utm_medium=cpc&ef_id=Wk1uqwAAAFDMc1LW:20180104041928:s> [↑](#endnote-ref-4)
5. ) <https://yourstory.com/2015/10/gapoon/> [↑](#endnote-ref-5)
6. ) <https://inc42.com/buzz/decoding-100-billion-home-services-marketplaces/> (Accessed on August 19, 2017) [↑](#endnote-ref-6)