Analysis on Customer Inquiry Resolution for Call Center

N.J.R Akila Shashith Karunasena

CDS\_TPFD/2018/01

This report is submitted as a partial fulfillment of Module 6 – Individual Mini-project

of the Training Programme Fundamentals of Data Science

conducted by the Center for Data Science

**a) Abstract**

Problem Statement:

One of the most important requirements for customer care center operation is providing timely solution to customer inquiries. Customer care center agent assign a committed date for every inquiry by customer when an inquiry is made. But there are occurrences when the issue is not resolved on agreed time due to various reasons. So, Identifying the factors affecting late issue resolving time is important for providing realistic committed dates in future.

What was done:

Created statistical model to predict customer inquiry is resolved on time or not.

Created BI dashboard to visualize the data.

How it was done:

In this scenario, there are two outcomes for every customer inquiry. Either it is resolved on time or not. So logistic regression method can be used to classify each customer inquiry into these two categories. Hence, I used logistic regression for this classification problem.

Dashboard created using Microsoft PowerBI and hosted on Power BI service (a hosting environment for Power BI reports.)

Significant results:

I able to create logistic regression model to classify the customer inquiry. Accuracy of the model is around 88%.

Recommendations:

I could use small proportion of data from whole dataset due to unavailability of both agreed date and issue closed date. If there were more usable data, I could create more accurate model.

I have only 1 month of data. If I able to access more data, the model might be more accurate.

And having data spanning through longer time can be used to create time series analysis. This type of analysis is useful to identify periodic patterns happening on over time.

**b) Table of contents**

Introduction ------------------------------------------------------------------------ 2

Methods and Tools ------------------------------------------------------------------------ 3

Analysis ----------------------------------------------------------------------- 4

Discussion and Conclusion ------------------------------------------------------------------------ 7

**c . Introduction**

Providing customer care is utmost important task of telecom providers. Providing good customer care is an important factor to attract customers to the network. On the other side, providing poor customer care is makes customer unsatisfied with the telecom provider and ultimately leave from the network.

Mostly customer contact the customer care center for some common issues. It can be for get an information about newly added feature or service. For these issues, customer care agents (here after called as agents) provides a solution on time. But some inquiries regarding technical problems and it need to be handled by technical staff. Sometimes these kinds of issues might take more time than estimated time.

So, Identifying the issues which take unusual time in the first place is useful to provide realistic resolve time to client. So, we are finding factors affecting to late issue fixing for customer inquiry.

**d. Methods and Tools**

Methods:

Since this is binary classification of client issues, classification method called Logistic Regression is used to do the classification.

Tools:

In here I used R (with R Studio) and some R packages available in CRAN to do the modeling.

Also, I used Microsoft PowerBI to data visualization

**e. Analysis**

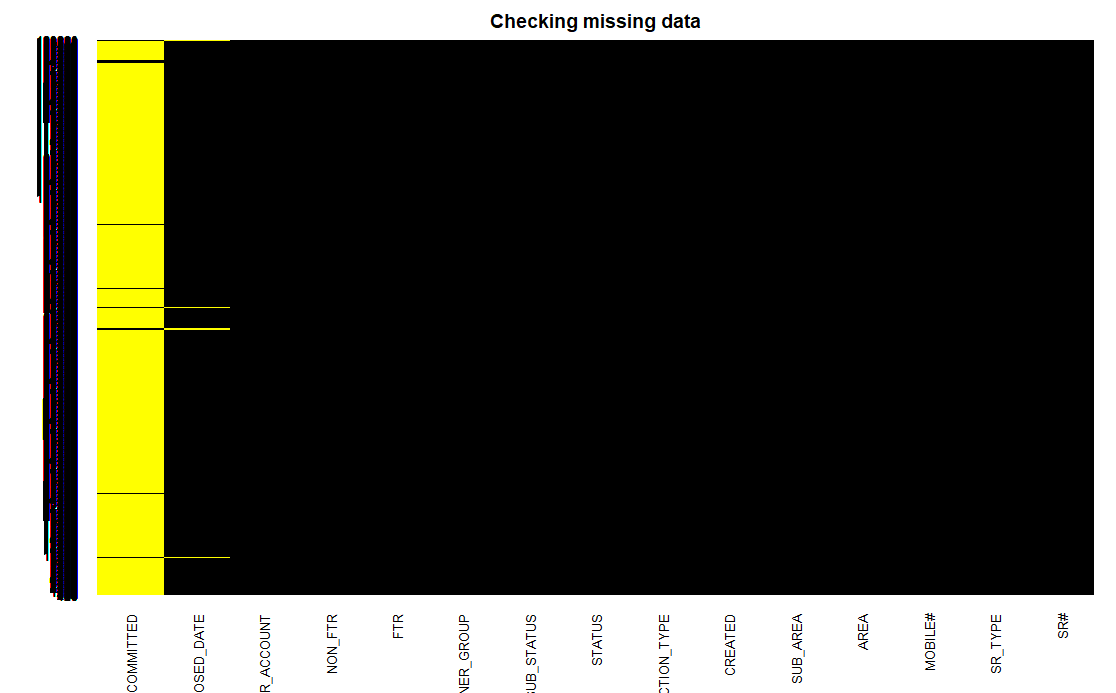
data set is provided in Excel file. Which had 139324 observations for Month of September. Which is the data for total inquiries for September month.

Let’s look at the structure of the data

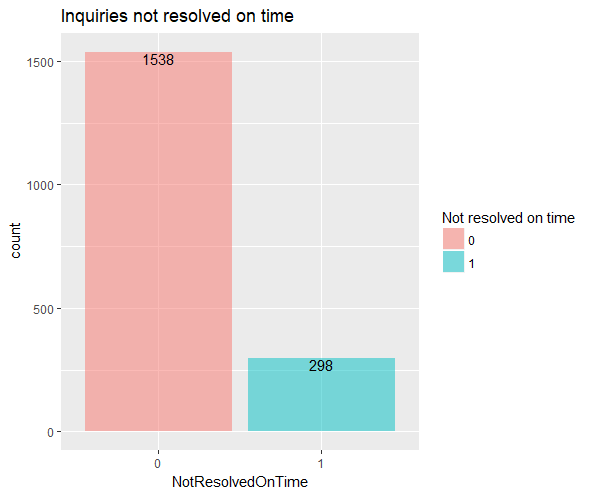
|  |  |  |  |
| --- | --- | --- | --- |
| SR# | chr | Service Request No | Request identification number |
| SR\_TYPE | chr | Service Request Type | Type of the request (Complain, request etc) |
| MOBILE# | chr | Mobile No | Customer's mobile no |
| AREA | chr | Related Area |  |
| SUB\_AREA | chr | Related Sub Area |  |
| CREATED | chr | Created Date |  |
| CONNECTION\_TYPE | chr | Type of connection |  |
| STATUS | chr | Status of the request | ongoing, closed etc |
| SUB\_STATUS | num | sub status of the main request |  |
| DATE\_COMMITTED | chr | Estimated deadline for request |  |
| OWNER\_GROUP | chr | Department who owns the issue |  |
| CLOSED\_DATE | chr | Actual closed data |  |

As per imported data most of the variables imported as a character.

Let’s look at how many null values exists in the dataset. For this I used R package called ‘Amelia’



There are lot of missing values in “DATE\_COMMITED’ variable and relatively few missing values in ‘CLOSED\_DATE’ variable. Both values important for our analysis. Therefore, I removed data with empty values for above date fields.

Only 1836 records having both ‘DATE\_COMMITED’ and ‘CLOSED\_DATE’ values available. From those records 1538 inquiries resolved on time and 298 inquiries not resolved on time.

Then I performed some techniques to anonymize sensitive data. I used ‘anonymize’ R package to encrypt mobile numbers and service request numbers. Also, I mapped a identification number for each ‘STATUS’,’Sub\_Status’ and ‘Owner Group’. In this analysis these identification number use to represent the underlying value.

Then the next step is conversion of data types. This is important step as most of imported data stored as a character value. I used R lubridate package for convert character into date time objects. And I get the day values of these variables into two separate columns and which will be using in the model creation.

After that I created a binary factor variable to identify whether the inquiry is resolved on time or not. This variable act as a dependent variable for the logistic regression model.

Variable “OwnerGroup” has 70+ levels. When I split the data into training and testing set some of the values not split between both the groups. As a example some levels only has one record. So I put these values into bins of size 10. And created new variable ‘OwnerGroupRanges’ to hold these values.

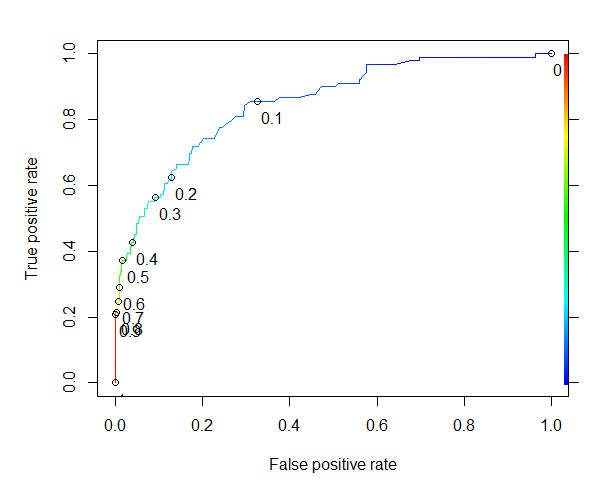
When I observed the data some of the variable can not use for the model. As a example ‘STATUS’ variable holds only 1 value for entire dataset. Likewise, there are few variables not usable for the model. Finally I used OwnerGroupRanges,Area,Closed\_day and committed\_day variables for the model.

**f**. **Discussion and Conclusion**

I created confusion matrix for model prediction vs actual test outcome, and it predicts the model accuracy of 83%.

According to model summary OwnerGroupRange3, AreaID13,AreaID17,AreaID4 to AreaID9 and Monday as closed day are negatively impact for the depending variable. That means above levels in variables **positively** strongly affect to resolve the inquiries on time.

I chose 0.2 as threshold value. Because I need to eliminate false positives as possible. After taking 0.2 as the threshold value the model predict the outcome of accuracy of 83%.



Sensitivity of the model = 55/(55+34) = 0.61

Specificity of the model = 402 / (402+59) = 0.87

Accuracy of the model = (55+402)/(55+402+59+34) = 0.83

Out of 139324 records only 1836 records having both committed date and closed date available. If there more complete data records available, it will be important for finding more insights from the data set.

Analyzed dataset collected through only for single month. If their data available spanning through multiple months data can be use to create time series forecasting. This is used to predict behavior of the trends of customer inquiries.