Customer Retention Case Study

Submitted by:

Deepro Sengupta

ACKNOWLEDGMENT

This project would not have come to fruition without the kind guidance of our senior/mentor Mr. Sajid Choudhary at FlipRobo Technologies. His help and input have been invaluable to this project. Secondly, I would like to extend my gratitude to all mentors and teaches at DataTrained for training me to be a good data scientist. Lastly, I would like to thank my parents for their support.

INTRODUCTION

E-commerce has gained a lot of traction in India and has become one of the large markets for global e-com giants. With more and more people gaining access to smartphones and internet in tier II and III cities as well as in villages, the ecom ecosystem in India will only grow. This has only made ecommerce a very competitive space. To add to this competition, any entrepreneur can have their ecommerce website up and running within a few hours with zero coding skills using services like Shopify and Woo-Commerce. Therefore, it is essential for existing market players to understand their existing customer base in order to retain them and also to improve their service and user experience.

This case study tries to understand the ecommerce customer in India.

METHODOLOGY

The problem we have in this analysis is that of understanding the customer base and customer retention in the context of Indian Ecommerce environment.

The objective, therefore, is to perform data analysis. A machine learning model is not required for this project.

In order to carry out the analysis, the data presented to us contained responses from a survey where respondents were asked various question ranging from their gender to the websites they use to shop online, etc. we primarily used hypothesis testing (Chi-square test of independence) and visualizations. Since no target column was specified, I started with first visualising each of the columns. Once done, I tried to draw insights from these visualisations which I presented on the presentation file.

Since all the columns were categorical columns, I also performed chi-square test on the data by first generating a list containing all combinations of columns using the following code:

```
columnCobination = [i for i in list(product(customerDF, customerDF)) if i[0] != i[1]]
```

I then iterated through this list to perform chi-square test on each of the pairs using the function defined below:

```
#Defining a function to do all of the above steps and tell us if the target column has any significant relation with other categ
def chi2IndependenceTest(col1, col2):

    crossTab = pd.crosstab(customerDf[col1], customerDf[col2], margins=True, margins_name='Total')
    observedVal = crossTab.values
    expectedVal = stats.chi2_contingency(crossTab)[3]
    rowNum = (crossTab.shape[0] - 1)
    columnNum = (crossTab.shape[1] - 1)
    doF=(rowNum-1)*(columnNum-1)
    chiSquare=sum([(o-e)**2./e for o,e in zip(observedVal, expectedVal)])
    chiSquareStatistic=chiSquare[0]+chiSquare[1]
    pValue=1-stats.chi2.cdf(x=chiSquareStatistic,df=doF)

    return doF, pValue
```

This project uses the Anaconda Distribution of Python and Jupyter Notebook for analysis. The system requirements are:

- License: Free use and redistribution under the terms of the ../eula.

- Operating system: Windows 8 or newer, 64-bit macOS 10.13+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.
- System architecture: Windows- 64-bit x86, 32-bit x86; MacOS- 64-bit x86; Linux- 64-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.

The following packages were used in this analysis:

1. NumPy: For basic mathematical operations
2. Pandas: For DataFrame manipulation
3. Matplotlib & Seaborn: For data visualization