
# ***customer activation and retention***

Submitted by

Nidhi Gupta (internship 16)

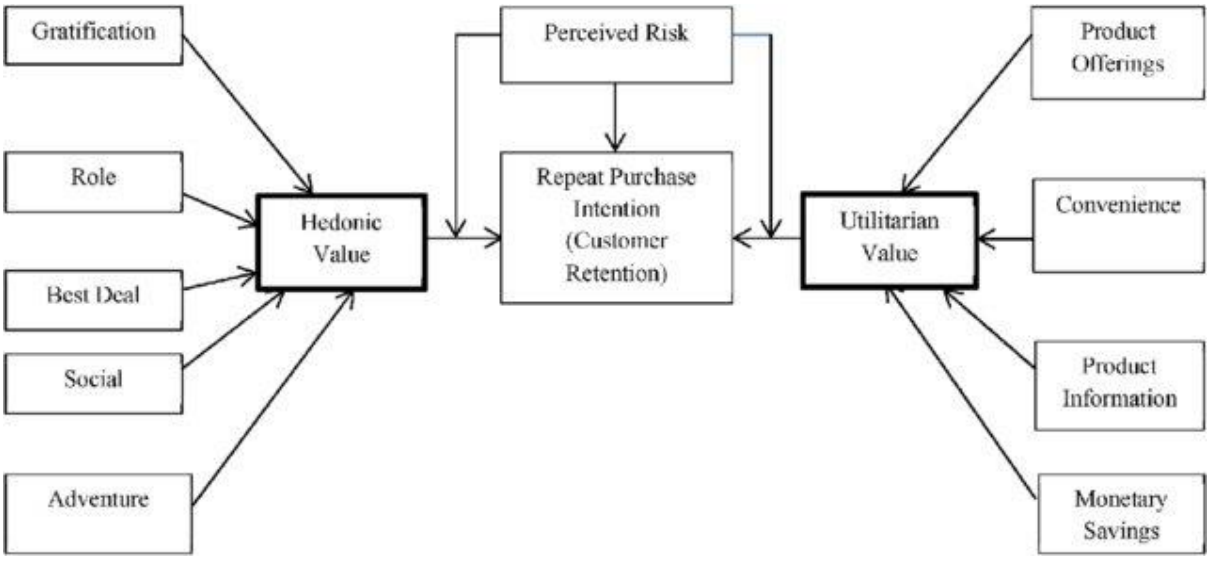
# INTRODUCTION

Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

# Problem analysis

Retention analysis (or survival analysis) is the process of analyzing user metrics to understand how and why customers churn. Retention analysis is key to gain insights on how to maintain a profitable customer base by improving retention and new user acquisition rates.

Overall, this analysis allows you to see how well your customer retention efforts are working. Without it, you may end up spending your marketing budget inefficiently




# ANALYTICAL PROBLEM FRAMING

First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.

- Drop duplicates rows if present in dataset.+Then we check for the null values present in our dataset.
- If null values are present then fill it via mean, median or mode.
- We also check the correlation of our dataset to check the correlation of the columns with each other. If columns are highly correlated with each other let's say 95% or above then remove those columns to avoid multi coli-nearity problem.
- We extract data from date column and make new columns like day, month and year to see the outcomes with our target column that is label.
- We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.
- We cannot remove outliers because more than 5-8% of our data are removed
- We have encoded all the columns for the further process.

# VISUALIZATION USED

- We plot correlation matrix via heatmap to see the correlation of the columns with other columns.
- We also visualize the correlation of columns with target column via strip plot to see which column is highly correlated with target column.
- We see the number of male and female customers with the help of count plot.
- We plot histogram to displays the shape and spread of continuous sample data.
- We also see the customers labels i.e male / female according to city and pincodes with count plot.
- We also visualize the correalation of columns with column via density plot , box plot to see correlation with target column.

# Modeling

We know that this is classification problem so we use accuracy score, classification report and confusion we have used test train model, decision tress classification, random forest classification, and search regression vector to find the best model .

Then we did cross vadidation for being more accurate and hyperparameter tuning.

# CONCLUSION

Our best model is Random forest classification with 98% accuracy, which is also seen through cross validation . hypertuning tells RandomForestClassifier(criterion='gini', max_depth= 8).