CUSTOMER RETENSION CASE STUDY

Problem statement:

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 definition:

Customer segmentation is a process where we divide the consumer base of the company into subgroups. We need to generate the subgroups by using some specific characteristics so that the company sells more products with less marketing expenditure. Before moving forward, we need to understand the basics, for example, what do I mean by customer base? What do I mean by segment? How do we generate the consumer subgroup? What are the characteristics that we consider while we are segmenting the consumers? Let's answers these questions one by one.

Basically, the consumer base of any company consists of two types of consumers:

1. Existing consumers
2. Potential consumers

Generally, we need to categorize our consumer base into subgroups. These subgroups are called segments. We need to create the groups in such a way that each subgroup of customers has some shared characteristics. Example

Suppose a company is selling baby products. Then, it needs to come up with a consumer segment (consumer subgroup) that includes the consumers who want to buy the baby products. We can build the first segment (subgroup) with the help of a simple criterion. We will include consumers who have one baby in their family and bought a baby product in the last month. Now, the company launches a baby product that is too costly or premium. In that case, we can further divide the first subgroup into monthly income and socio-economic status. Based on these new criteria, we can generate the second subgroup of consumers. The company will target the consumers of the second subgroup for the costly and premium products, and for general products, the company will target consumers who are part of the first subgroup.

When we have different segments, we can design a customized marketing strategy as well as customized products that suit the customer of the particular segment. This segment-wise marketing will help the company sell more products with lower marketing expenses. Thus, the company will make more profit. This is the main reason why companies use customer segmentation analysis nowadays. Customer segmentation is used among other domain such as the retail domain, finance domain, and in customer relationship management (CRM)-based products.

Data Analysis:

The key to success in an organisation is the ability to attract and retain top talents. It is vital for the Customer Retention case study project to identify the factors that keep employees and those who the first stage of this analysis is to describe the dataset, understand the meaning of each variable, detect possible patterns and perform the necessary adjustments to ensure that the data will be proceeded correctly during the Machine Learning process. each prompt them to leave. Organisations could do more to prevent the loss of good people.

Data preparation cleaning:

* Reading the CSV file and doing initial statistical analysis (shape, values etc)
* Data Pre-processing: Reading the uniques values for each column and removing those which won’t be significant in the analysis further.
* Create a new data frame to proceed with the analysis further

EDA Concluding Remarks:

* Find patterns of data through visualization and reveal the hidden trends from data.
* Using both matplotlib and seaborn library to visualize the data
* Finding relationships between features using bar graphs, histograms, box plots, heatmap
* Analyzing both the numerical and the categorical columns separately

Pre-processing pipeline:

For the model to proceed with the data efficiently, the categorical variables salary and department have been encoded. As the values of salary have an order, they have been encoded into integers within the same variable. For department, as the values have no specific order, they have been encoded into individual variables with Boolean values. Thus, the dataset has been transformed from 10 variables to 19 variables. Numerical variables scaled between 0 and 1 to remove any influence of their difference in value ranges on the model. They have also been checked for skewness, without a real change on their shape.

Building machine learning model:

As the dataset is imbalance, use cross validation when training the models, and each baseline model performance can be tabulated.

The model will be cross-validated using a 10-fold cross validation method returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the training set split

Here we using the classification Methods to build the models:

There are 6 classification models

1. DecisionTreeClassifier

2 Logistics Regression

3 KNeighborsClassifier

4 SVM Classifier

5 Navie byes Classifier

6 Random Forest Classifier

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Logistic Regression has achieved F1-score 100.0% and **Recall 100.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the DecisionTreeClassifier has achieved F1-score 100.0% and **Recall 100.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the KNeighborsClassifier has achieved F1-score 90.0% and **Recall 88.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the SVM Classifier

has achieved F1-score 100.0% and **Recall 100.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Navie byes Classifier has achieved F1-score 93.0% and **Recall 96.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Random Forest classifi

r has achieved F1-score 100.0% and **Recall 100.0%**.

Also, if features are closely related to one another .one of them has to be removed to prevent misleading results to linear models such as Logistic Regression. Although tree-based models are not directed affected, they could also lead to over-fitting.

According to the classification report the accuracy of the model is 87% however its recall is lower at 43% of positive cases. The RandomForestClassifiermodel is providing excellent results, however the purpose of the problem is to identify employees that are likely to leave. This is the reason that recall then becomes a very important measure. Recall measures the fraction of values that are identified correctly.

Random Forest Classifier has emerged as the final winning model with F1-score 100.0% and highest **Recall 100.0%**. This could be the highest possible score achieved with the inherent limitations in the dataset.

The top factor for employee attrition in this hypothetical organisation seems to be **monetary**, emerged at the top. This could be due to a bad compensation process or causing a poor work-life balance. The next important factor seems to be **personal relationships**with follow workers, where current manager and job role could be the main contributing reasons for attrition. Finally, **employee engagement**is a critical satisfaction factor, and the organisation should keep employees constantly involved and motivated

Machine learning models are as good as the data to feed it, and more data would strengthen the model. For example, in this dataset, the feature ‘Performance Rating’ has been restricted to scores of 3 and 4 only. More insights could be generated if the full spectrum of performance ratings is included. In the real-life situation, getting the right data is often more challenging than the analytics itself.

Concluding Remarks:

Customer retention case study is gaining traction in organisations that embrace digital transformation. The scope has expanded from analytics of employee work performance to providing insights so that decisive improvements can be made to organisational processes. While some level of attrition is inevitable, it should be kept at the minimal possible level.

This model will allow the company to calculate the probability of an employee to leave the company and to act on key-factors to avoid departures. The satisfaction of employees and the amount of workload they have to bear seem to be important causes of withdrawals. A particular attention on the work-life balance would be crucial to improve the turnover rate.