Walmart trip type classification problem, to classify shopping trips.

Abstract:-

In this fast changing world e-commerce platforms face cutting edge competition, so in order to optimize their shopping process, Walmart uses trip type classification to segment its shoppers and their store visits to improve the shopping experience for customers. This trip type uses customer's shopping history to find their shopping pattern and by using the customer insights to classify trip types. Trip type classification takes advantage of customers' behavioral approach,psychological approach and geographic approach in market selection and market segmentation. In order to achieve this trip type classification task we use Data analysis and Machine learning techniques to classify customers purchase data into different trip types. The goal of this classification is to refine the trip classification process.

1- Problem definition:-

The task is to categorize shopping trips of the customer by using insights from historical customers shopping data i.e based on the items customers purchased.

The purpose of this project is to improve customer segmentation enabling personalized services and targeted advertising.

This task will help the service provider in understanding customer interest i.e the products in which customers are more interested, and by understanding this the service provider will be able to better equip itself to fulfill the required customers demand.

So, in the real world this task will help in enhancing relation among the customers and service provider.

2- Dataset:-

As self acquisition of a valid real world dataset for analysis is very difficult, so this project takes the dataset from Kaggle, which is good and reliable source for obtaining dataset for analysis

Link for the dataset:-
https://www.kaggle.com/competitions/walmart-recruiting-trip-type-classification/data

This dataset has 7 different types of fields each explaining different aspects of the dataset, which are as follows:-

TripType:- In this dataset different shopping trips i.e made by the customers are categorized into 38 distinct trip types, each trip type is a collection of different trips made by the customers, these trip types are indicated in numeric values.

visitNumber:- It is an id number assigned to every trip i.e made by the customer.

weekday:- This field shows that on which particular day in a week the trip is made by the customer.

Upc:- This is the **unique purchase code** that is provided to each product i.e purchased or returned by the customers.

ScanCount:- It shows the counts or number of items of the same category i.e purchased or returned by the customers. Positive value of ScanCount shows that the item is purchased by the customer and its negative value shows the customer has returned the purchased product.

DepartmentDescription:- This field tells the domain or the category of the items purchased by the customers in each trip.

FinelineNumber:- It is also a unique categorical number which is used to categorize different items or products.

This Kaggle has a huge dataset, it comprises 647054 rows and 7 fields or columns. It is a multiclass classification problem(a classification task with more than 2 classes), because we have to categorize shopping trips into 38 distinct trip types. To solve a multiclass classification ML techniques like Decision tree, Random forest, XG boost, K-NN can be used.

The is an imbalanced dataset, because in this classification problem classes are not represented equally i.e here there is a huge variation in counts of all the trips types.In order to counter this challenge oversampling and undersampling can be used, but in case where dataset is highly imbalance, we can carry our analysis by just ignoring the outlier.

To analyze and process the data tools like Python, python libraries(NumPy, Pandas, seaborn, matplotlib, sklearn etc.) can be used.

3- Key metrics(KPI) to optimize:-

Business Metric:- A business metric is a business measure used to monitor and assess the success and failure of various processes used in a business. Business metrics is used to evaluate the company's progress and performance in long term and short term goals, within the estimated time frame.

Important metrics to estimate performance:-

KPI-1 Most popular product department in TripType:-

This project shows the correlation between fields TripType and DepartmentDescription to optimize the count to show maximum value of department Description i.e. it shows max count of products or items purchased in a particular triptype and is grouped by TripType.

This analysis is very useful according to the business perspective, because it directly shows the trip type with the corresponding product purchased in the Department and their maximum count value.

This tabular representation of the used fields helps to understand the customers perspective i.e likes and dislikes of customers. It gives the shopping pattern of the customers by using their shopping history i.e it shows which are the highest selling products in each trip types, this helps the service provider to understand the needs and demand of customers by looking at most and least selling products in each trip type.

By understanding the shopping pattern and interests of the customers, the company can provide better schemes and offers to the customers, which in turn will benefit the customers and service providers will be able to provide a better service to the customers.

Service providing organizations will be able to look upon those products and services that are least attractive for the customers in every trip type.This will help the organizations to understand the low demand products and replace them with the in demand products. Although it may result in discontent among the few people who are

purchasing the low demand products, overall it will be beneficial for the majority of the people as well as for the service providing organizations.

This analysis will help the organizations to cater better services to a wide range of groups in the specific region, because in different geographical locations people have different needs and demands, so it's not necessary that the product or service that is popular in a particular location will also be popular in other locations too.

By using this performance metric the service providing organizations can increase their performance by providing better services and products to the majority of customers in the region. It will also build a healthy relation between the customer and service providers.

	TripType	max_category_count	DepartmentDescription	count_values
0	3	5369	FINANCIAL SERVICES	5369
1	4	563	PHARMACY OTC	563
2	5	5281	PHARMACY OTC	5281
3	6	1000	LIQUOR,WINE,BEER	1000
4	7	3669	SERVICE DELI	3669
5	8	4747	DSD GROCERY	4747
6	9	1383	MENS WEAR	1383
7	12	278	DSD GROCERY	278
8	12	278	HOUSEHOLD PAPER GOODS	278
9	14	15	FABRICS AND CRAFTS	15
10	15	2294	CELEBRATION	2294
11	18	1160	TOYS	1160
12	19	491	ELECTRONICS	491
13	20	1639	AUTOMOTIVE	1639
14	21	1283	FABRICS AND CRAFTS	1283
15	22	643	ELECTRONICS	643
16	23	112	PLAYERS AND ELECTRONICS	112
17	24	2212	COOK AND DINE	2212
18	25	5381	MENS WEAR	5381
19	26	758	HARDWARE	758
20	27	1325	LAWN AND GARDEN	1325
21	28	1250	SPORTING GOODS	1250
22	29	480	TOYS	480
23	30	1140	SHOES	1140
24	31	989	WIRELESS	989
25	32	4713	INFANT CONSUMABLE HARDLINES	4713
26	33	4067	HOUSEHOLD CHEMICALS/SUPP	4067
27	34	2657	PETS AND SUPPLIES	2657
28	35	6135	DSD GROCERY	6135
29	36	8635	PERSONAL CARE	8635
30	37	13351	PRODUCE	13351
31	38	7252	DAIRY	7252

	TripType	max_category_count	DepartmentDescription	count_values
32	39	12956	DSD GROCERY	12956
33	40	32639	GROCERY DRY GOODS	32639
34	41	340	SHOES	340
35	42	1278	IMPULSE MERCHANDISE	1278
36	43	523	PERSONAL CARE	523
37	44	1857	PERSONAL CARE	1857
38	999	2289	FINANCIAL SERVICES	2289

Above shows the Table of KPI-1

KPI-2 Metrics to evaluate the performance of the multiclass classification model:-

Precision:- It is the accuracy of the positive predictions.
tp = true positive; fp = false positive

Precision = tp / (tp + fp)

Recall:- It gives the fraction of positives correctly identified among total tp and fn.
fn = false negative

Recall = tp / (tp + fn)

F1 score:- It takes both Precision and recall into account. F1 score is the weighted average of Precision and Recall.

F1 score = 2(Precision * Recall) / (Precision + Recall)

For a Balanced dataset we use accuracy as metric, but this does not work for an imbalanced dataset. In case of an imbalanced dataset Precision, Recall and F beta score (F1 score, for Beta = 1) can be used as evaluation metrics to validate our model.

In cases where we want to reduce False negative, Recall is the best evaluation metric to use.

Precision indicates the quality of positive prediction made by the model. It is used where the correctness of the model is of prime concern.

In cases where we want to maximize both Precision and Recall F1 score can be used.If we want our model to be both correct and not to miss any correct prediction this is the evaluation metric that can be used.

Implementation of metric KPI-2 in Python using NumPy and Pandas is shown below:-

In [7]:

```
import pandas as pd   # importing python modules
import numpy as np
```

In [8]:

```
actual=pd.Series([1,1,1,1,1,0,0,0,0,1,1,1,0,0],name='Actual')

# 'pd.series' gives a one-dimensional labeled array,
# which is capable of holding data of any type.
# The axis labels are collectively called index.
```

In [19]:

```
pred=pd.Series([1,1,1,1,1,0,0,0,0,0,0,0,1,1],name='Predicted')
```

In [10]:

```
confusion_matrix=pd.crosstab(pred,actual)

# confusion matrix is a 2-D matrix,
# it is used to evaluate the performance of the classifier.

# it is used to evaluate the performance of a classifier by comparing,
# it's predictions to the real world values.

confusion_matrix
```

Out[10]:

Actual	0	1
Predicted		
0	4	3
1	2	5

In [16]:

```
a=confusion_matrix.to_numpy()

# converting a pandas dataframe of 2-D data structure,
# to a numpy array which represent the values in given index.
```

In [22]:

```
tp=a[1][1]
tn=a[0][0]
fp=a[1][0]
fn=a[0][1]

precision = tp/(tp+fp)
recall = tp/(tp+fn)
f1 = 2*(precision*recall)/(precision+recall)
print('Precision:-',precision)
print('Recall:-',recall)
print('F1 score:-',f1)
```

```
Precision:- 0.7142857142857143
Recall:- 0.625
F1 score:- 0.6666666666666666
```

4- Real world challenges:-

- The dataset is imbalanced, so to solve this we generally use overfitting and underfitting, but for highly imbalanced dataset this solution also fails.

  For highly imbalanced dataset we will have to remove the outlier points on dataset and must continue the required analysis by using the remaining points.

- We don't know whether the dataset is statistically legit or not, because we don't know if we will get same result with similar problem using different dataset, as in different geographical locations customers interests and requirements vary, so it is not necessary that the service provider can provide same services in different locations.

  To deal with this problem, organization must have experts for data collection, so they can collect legit data that covers wide spectrum of customers spread is different geograplical locations, by this we will have better correctness of result for customers of different locations.

5- References:-

- https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

- https://www.datarobot.com/blog/multiclass-classification-in-machine-learning/

- https://www.kaggle.com/competitions/walmart-recruiting-trip-type-classification/data
- https://github.com/shockwave22/Walmart-Trip-Type-Classification
- https://datawookie.dev/blog/2016/01/kaggle-walmart-trip-type-classification/
- https://www.youtube.com/watch?v=mhI7hzVRr-k&t=120s

.