THE PROBLEM STATEMENT FOR CASE STUDY

A fast-growing fintech startup has built a user base by providing a payment platform for users. They are now looking to build a credit product and make it available for their customer base. Credit products comes with huge risks, and they want Data Science team to help them with a model/solution which will enable them to do risk profiling of potential customers for the credit product.

Firstly we will Crunch down the Banking term to make it understand to our stockholders.

What are Credit Products ?

Credit Product means a charge or credit card or program, an on-line or mobile credit or charge account, or other credit or charge device. For avoidance of doubt, this term does not include debit only products.

What is Risk Profiling ?

Risk profiling is the evaluation of an individual's willingness and ability to take risk. Risk profile is very subjective and may vary based on your age, financial circumstances, investment objective, personal experiences, risk-return expectations and time horizon to achieve the goals.

And their 5 types are follows :

1.Conservative

2.Moderately Conservative

3.Moderate

4.Moderately Aggressive

5.Aggressive

1. How would you approach this problem?

A. Gathering the Data from various sources for model building.

B. Data preprocessing, Noise and treating the Outliers for model fitting.

C. Using various Data Visualization techniques and reporting to see the trends in customer behavior.

D. Performing feature engineering mechanism and algorithms to select important variable based on weightage.

E. We will use Machine Learning Algorithms to identify key variables from number of variables.

D. Machine Learning algorithms for customer segmentation based on certain variables and come up with strategy for each customer segment.

2.From where would you gather the data?

A. We can build a questionnaire to be filled by the visiting customer for payment in order to get the data form the customers.

B. Using questionnaire data are more easy to handle by the machine learning algorithms as they do not contain huge amount of Noise.

C . Although we can use data from open-source for our early stages for risk profiling.

D. API also known as web hooks can also play an important role for collecting the data in our case.

3.How would you clean and explore the data?

A. Converting Categorical or ordinal Data into Numeric Data

Handling Non-Numeric Variables we will select all the types of object type(Non Numeric)

A. For Categorical Variables (Do not have Ranking) = We will create Dummies using (get_dummies)

B. For Ordinal Variables (They Have Order)= We will use Map function.

B. Handling Missing Values

It is very important to treat missing values properly when we are dealing with the sensitive data of the customer.

A. We can use fillna() method for replacing missing values with mean , Median or most frequent value.

B. We can also use Iterative Imputer will consider the missing variable to be the dependent variable and all the other feature will be independent variables.

C. Detecting and Treating the Outlier

We can use various methods for outlier treatment such as

1.Visual methods to spot and remove outliers we can use

i. Box-plot

ii. Scatter plots

2. We can also use mathematical function mentioned below

i. DBScan clustering

ii. Isolation Forest =Isolation Forest method can handle missing values well and does not require scaling of inputs.

EXPLORATORY DATA ANALYSIS

We can create various visualization for the basic understanding of data

i. We can create Histogram on income group of the customer

ii. We can create a Sandhurst chart for the age bracket of the customer and also on nature of occupation of the customer if we have.

iii. Depending upon the features we have we can use numerous number of visualization for understanding our data better.

4.What features would you build in solution/model that will help business identify risk free potential customers?

A. We will be using Customer Segmentation for our problem.This will be done on the criteria provided for potential customer on risk profiling.

B. Machine learning algorithm that is suitable for customer segmentation problems is the k-means clustering algorithm for small or initial Data.

C. Although we can use other clustering algorithms as well such as DBSCAN, Agglomerative Clustering, etc.

D. For the huge amount or larger datasets we can use Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH).

E. It is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible.

EASE OF RETRANING

A. Customer Segmentation is not a “develop once and use forever” type of project.

B. Data is ever-changing, trends oscillate, everything keeps changing after model is deployed.

C. As we know more labeled data becomes available after development, and it’s a great resource for improving the overall performance of our model.

D. There are many ways to update customer segmentation models, but here are the two main approaches:

i.Use the old model as the starting point and retrain it.

ii.Keep the existing model and combine its output with a new model.

BETTER SCALING

A. Although we can use AI for same problem to be solved but using machine learning models deployed in production support scalability.

B. With the help of cloud infrastructure, these models are quite flexible for future changes and feedback as they have inherent capability to handle more data and scale in production.

HIGH ACCURACY

A. The value of an optimal number of clusters for given customer data is easy to find using machine learning methods like the elbow method.

B. Not only the optimal number of clusters but also the performance of the model is far better when we use machine learning.

Why are we using Clustering Algorithms ?

A clustering machine learning algorithm is an unsupervised machine learning algorithm. It's used for discovering natural groupings or patterns in the dataset.

Some of the most popular clustering algorithms are:

1.K-Means Clustering

2.Agglomerative Hierarchical Clustering

3.Expectation-Maximization (EM) Clustering

4.Density-Based Spatial Clustering

5.Mean-Shift Clustering

Why are we using K-Means Clustering for our problem ?

A. The algorithm discovers groups (cluster) in the data, where the number of clusters is represented by the K value.

B. The algorithm acts iteratively to assign each input data to one of K clusters, as per the features provided.

C. All of this makes k-means quite suitable for the search of groups of potential customer.

i. Finding optimal number of clustering

A. We have a few methods, such as the elbow method, gap statistic method, and average silhouette method, to assess the optimal number of clusters.

ii. For implementing the elbow method it takes two values as input:

* K (number of clusters),

* Data (input data).

D. The stage at this number of clusters is called the elbow of the clustering model and in our case it is K =5.(as we have the category as

1.Conservative

2.Moderately Conservative

3.Moderate

4.Moderately Aggressive

5.Aggressive)

FOR FUTURE DEVELOPMENTS :-

When the develop ML models are deployed we can run experiments as below

Those experiments may:

A. Using different models and model hyperparameters.

B. Using different training or evaluation data.

C. Running the same code in a different environment.

And as a result, they can produce completely different evaluation metrics.