Solution Design

* Background

Each and every retail company wants to understand the customer purchase behavior against various product categories for the Black Friday Sale The dataset is a sample of 550000 observations about the Black Friday Sale in a retail store and contains different kinds of variables either numerical or categorical. It also contains missing values.

* Problem Statement

The store wants to determine the customer purchase behavior for different products.It’s a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

* Objective

We would be building a model to understand the customer purchase behavior in order for these retail companies to create personalized offer against each and every product

**Step 1**

We have categorized our system as:

* Supervised Learning task**:** The training data is labelled i.e. we already know how much a customer spent on a specific product
* Regression task**:** The algorithm is expected to predict the purchase amount a client is expected to spend on this day.

**Step 2**

Selected a Performance Measure-Basically performed a few techniques in predicting the error the system makes in its predictions.

**Step 3**

Made assumptions

Looked at all the possible parameters that might influence the amount a customer spends on Black Friday.

1.City level Analysis

City type & Size, Population Density, Younger Population

2.Customer level Analysis

Income, Age & Gender, Family Size & Purchase History

3.Store Level Analysis

Location, Size of Store, Competition between stores

4.Product Level Analysis

Category, Price, Advertising, Visibility and Brand

**Step 4**

Analyzed the Predictors and Outcomes-

Gender, Age ,City\_Category, Product\_Category \_1 would be the predictors as they would be influencing more the amount spent by a customer.

Purchase amount is the target variable i.e. the Outcome.

**Step 5**

Have described the data as follows-

User\_ID -User

Product\_ID -Id Product

Gender -Boolean

Age -Age of customer

OccupationId -Occupation of each customer

City\_Category

Stay\_In\_Current\_City\_Years

Marital\_Status

Product\_Category\_1

Product\_Category\_2

Product\_Category\_3

Purchase -Purchase amount in dollars

**Step 6**

Data Analysis techniques-

• Linear Regression Model

The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variable(s).

Here we use multivariate linear regression to find the mathematical relation between the Purchase amount spent by the client as a function of ,Gender, Age, Occupation, City, Stay\_In\_Current\_City\_Years, ,Marital\_Status, Product Category .

1: Create the training and test data

2: Fit the model on training data and predict on test data

3: Review diagnostic measures.

4: Calculate prediction accuracy and error rates

5: Perform k- Fold Cross validation.

• Ridge Regression Model

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

We first implement it on our problem and check our results that whether it performs better than our linear regression model by observing the improvement in the value of R-Square.

The outcome is typically a model that fits the training data less well but generalizes better because it is less sensitive to extreme variance in the data such as outliers.

• Decision Tree Model

A Decision Tree is a supervised learning predictive model that uses a set of binary rules to calculate a target value.

1. Use recursive binary splititng to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.

Recursive Binary Splitting is a greedy and top-down algorithm used to minimize the Residual Sum of Squares (RSS), an error measure also used in linear regression settings.

2. Next, you apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α (an additional tuning parameter) that balances the depth of the tree and its goodness of fit to the training data.

We will use K-fold cross-validation to choose α.This technique simply involves dividing the training observations into K folds to estimate the test error rate of the subtrees. Our goal is to select the one that leads to the lowest error rate.

• Random Forrest Model

Random Forests is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job.

1. First assume that the number of cases in the training set is K. Then, take a random sample of these K cases, and then use this sample as the training set for growing the tree.

2. If there are p input variables, specify a number m<p such that at each node, you can select m random variables out of the p. The best split on these m is used to split the node.

3. Each tree is subsequently grown to the largest extent possible and no pruning is needed.

4. Finally, aggregate the predictions of the target trees to predict new data.

• XGBoost (Extreme Gradient Boosting)

It has both linear model solver and tree learning algorithms. So, what makes it fast is its capacity to do parallel computation on a single machine.

1. Convert the training and testing sets into DMatrixes: DMatrix is the recommended class in xgboost.

2. Specify cross-validation method and number of folds. Also enable parallel computation

3. Train your model.

4. Evaluate the model and plot actual vs predicted.