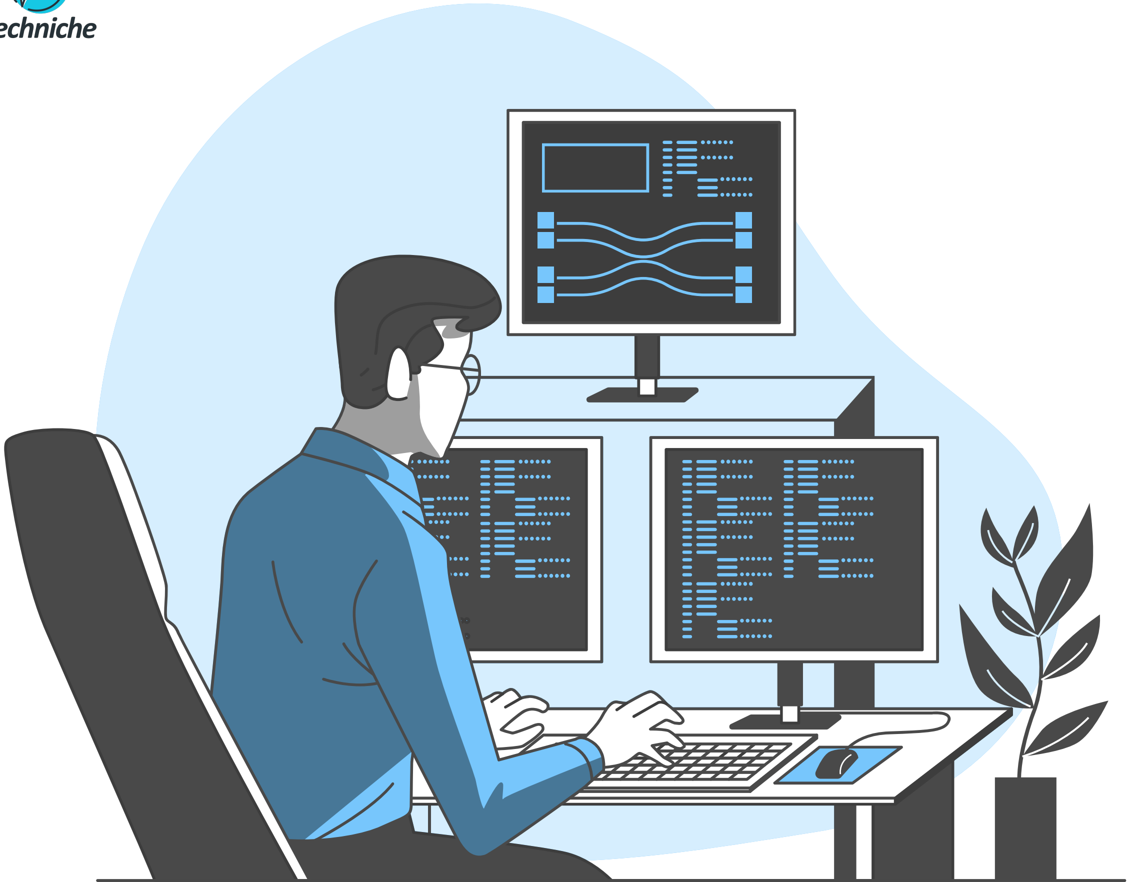
<- Click on this to start audio

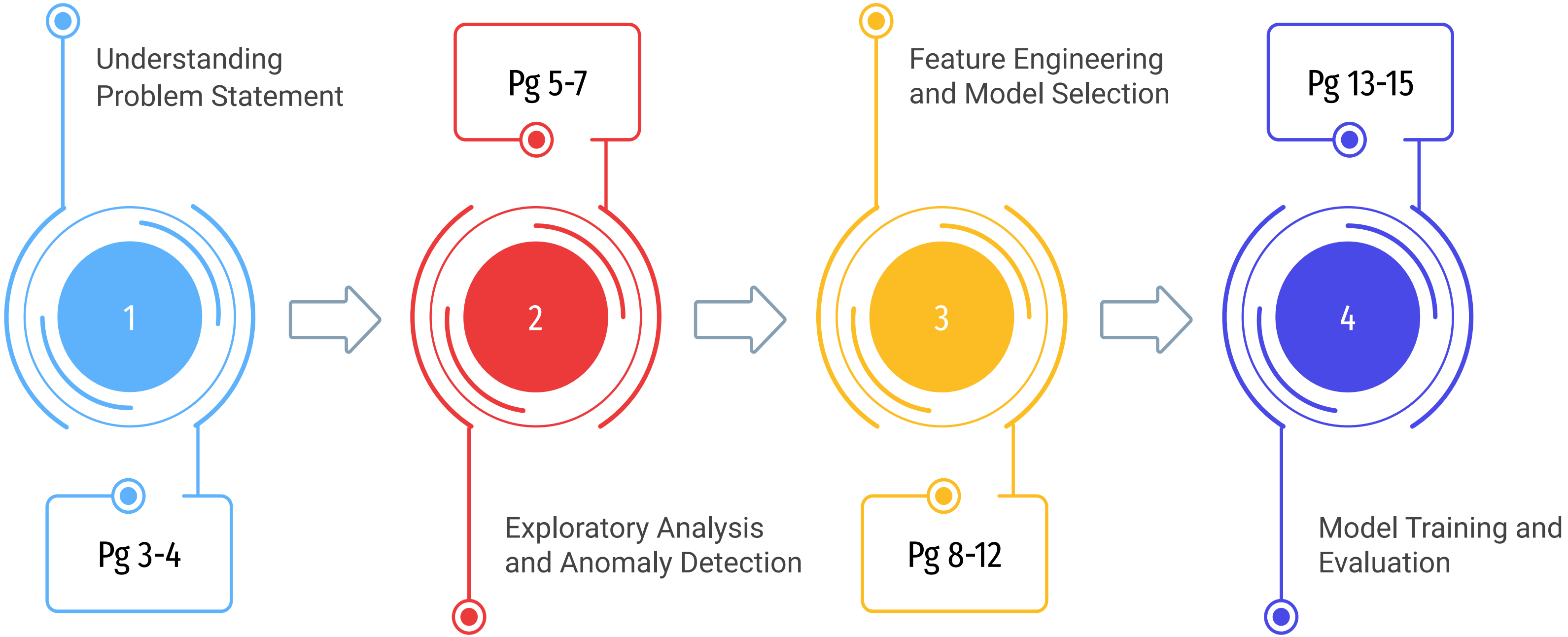


BEYOND ANALYSIS

TEAM NAME: OddAsparagus11

TEAM MEMBERS: VENMANI A D

Table of Contents



Understanding Problem Statement Pg 3-4

Exploratory Analysis and Anomaly Detection Pg 5-7

Feature Engineering and Model Selection Pg 8-12

Model Training and Evaluation Pg 13-15

Problem Statement

KEY OBJECTIVE:
Estimating the Customer Value and Temporal Component of the Customer Value for each customer using the given features

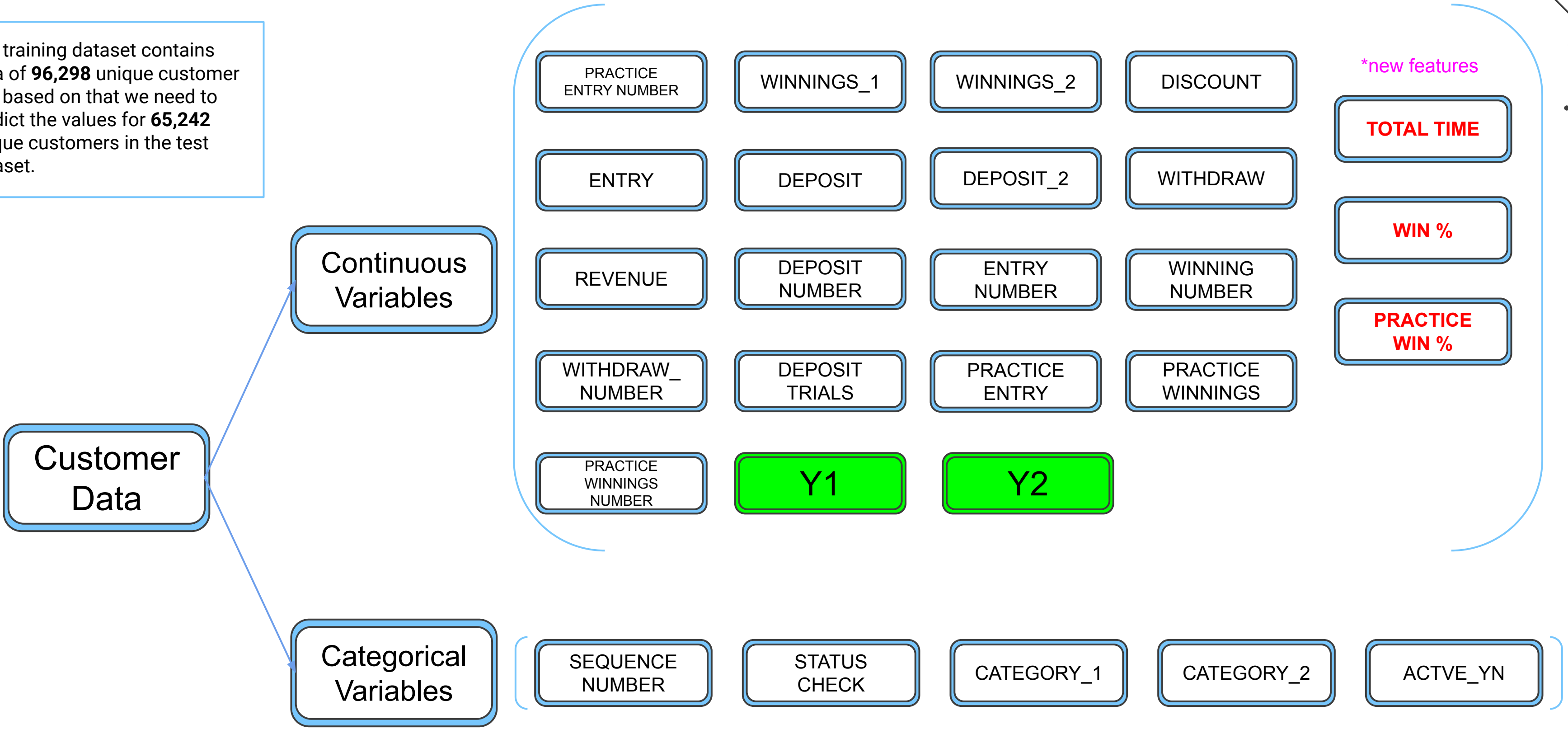
MORE SPECIFIC MODEL OBJECTIVE:
Achieving lower **Root Mean Squared Value** in the model for predicting Y1 and Y2.

BACKGROUND INFORMATION:

- Head Digital Works is now a well-funded and supported growth stage 'start-up' and prides itself as being one of the few consistently profitable enterprises in its industry
- Their main products include online monetary games like A23.COM / FANFIGHT.COM and their latest product is a Data Science Product named CRICKET.COM
- They wanted to predict how much each of their customers to the customer (monetarily) during their lifetime or in a certain amount of period
- Y1 and Y2 are the past values of worth of different customers. With those values and features we were asked to predict it for the customers in the test dataset

Understanding Data

The training dataset contains data of **96,298** unique customer and based on that we need to predict the values for **65,242** unique customers in the test dataset.



Customer Data

Continuous Variables

PRACTICE ENTRY NUMBER

WINNINGS_1

WINNINGS_2

DISCOUNT

ENTRY

DEPOSIT

DEPOSIT_2

WITHDRAW

REVENUE

DEPOSIT NUMBER

ENTRY NUMBER

WINNING NUMBER

WITHDRAW_ NUMBER

DEPOSIT TRIALS

PRACTICE ENTRY

PRACTICE WINNINGS

PRACTICE WINNINGS NUMBER

Y1

Y2

*new features

TOTAL TIME

WIN %

PRACTICE WIN %

Categorical Variables

SEQUENCE NUMBER

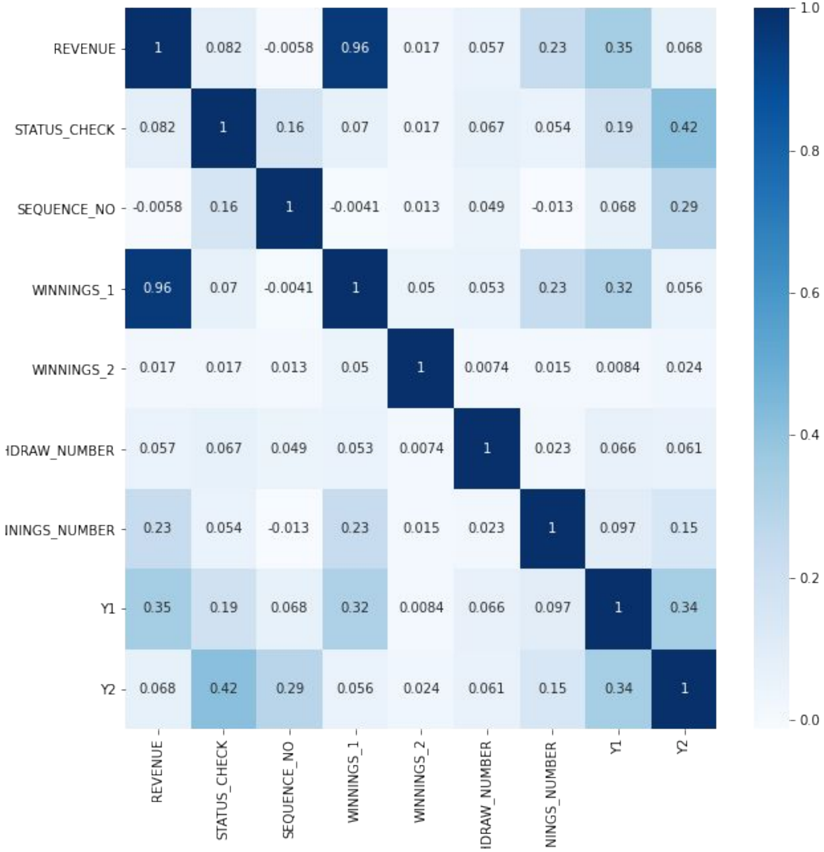
STATUS CHECK

CATEGORY_1

CATEGORY_2

ACTVE_YN

Exploratory Data Analysis



Top Correlated Variables for Y1:
Revenue
Deposit
Entry
Winnings_1
Status_check

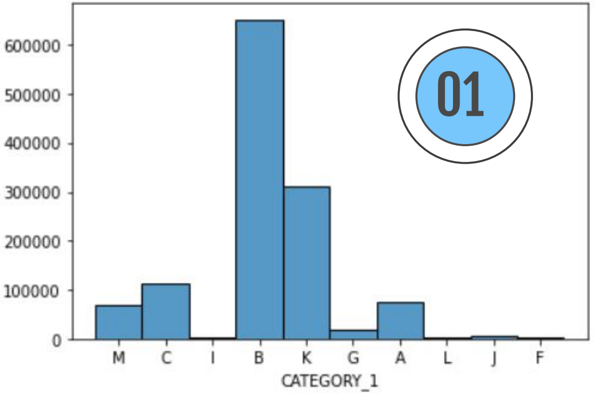
Top Correlated Variables for Y2:
Sequence Number
Status Check
Entry Number
Winnings Number

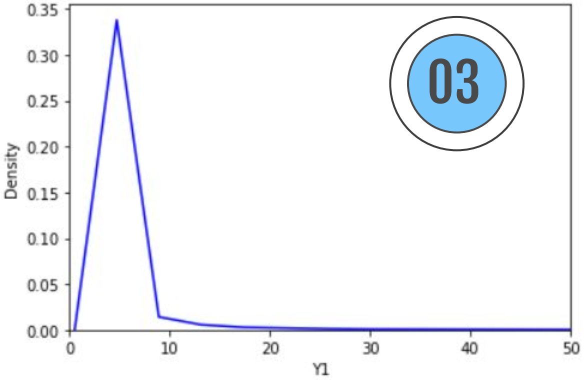
As the sequence number seems to be an important factor, I created a new feature which will denote the number of sequences for each customer. More the sequences means longer the customer has stayed with the company.

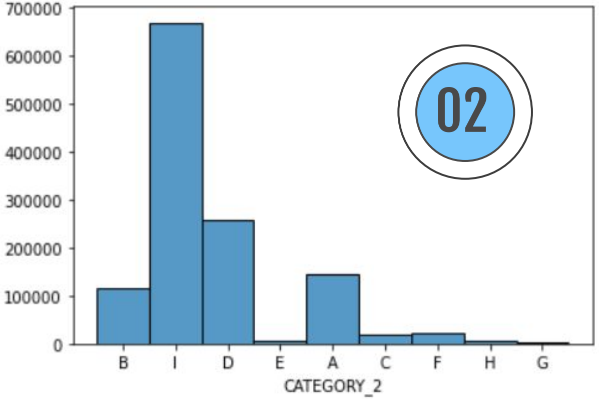
And also as the entry number and winnings number seem to be important, I created a new variable called percentage of wins.

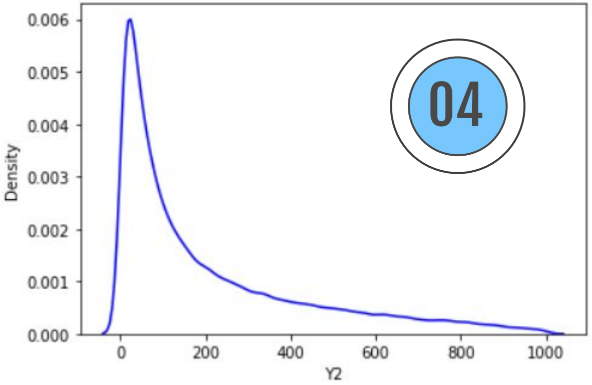
more about these features are discussed in the feature addition section

1 and 2 – Distribution of CATEGORY_1 and CATEGORY_2
3 and 4 – Density Plot of Y1 and Y2



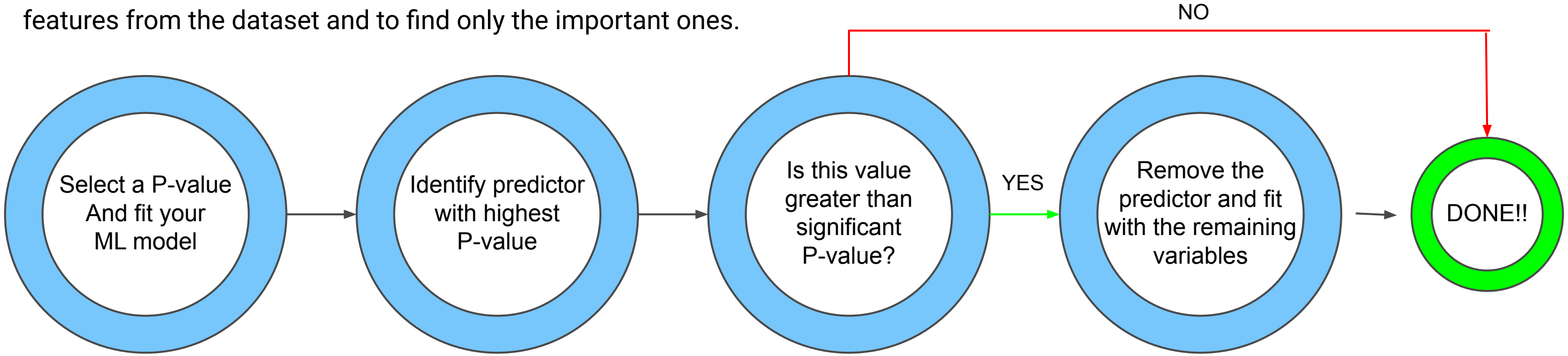






FEATURE SELECTION

Tried Backward Feature Selection to remove least relevant features from the dataset and to find only the important ones.


NO
Select a P-value And fit your ML model
Identify predictor with highest P-value
Is this value greater than significant P-value?
YES
Remove the predictor and fit with the remaining variables
DONE!!

ANOMALY DETECTION

- I noticed in the dataset, that some columns had anomalies or outliers in them
- Used **Isolation Forest** to name the worser 10% of the dataset as anomalies and labelled them into a separate column in binary manner - **-1 if it is not a outlier** and **1 if it is a outlier**
- Predicted the outliers in test dataset by using the model fitted on the training data and added a new column in the test data
- But this column didn't have much impact on the final results

Feature Addition

01

Active Period

Number of sequences for each customer in which they are active

02

Total Time

Number of sequences present for each customer in dataset

03

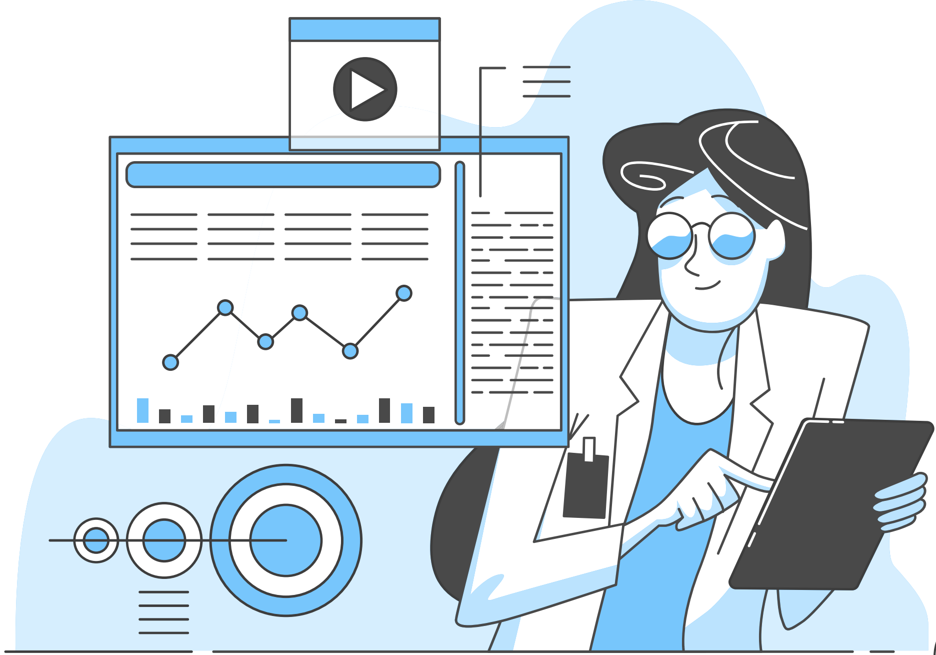
Win Percentage

Sum of winning number divided by sum of entry number for each customer

04

Practice Percentage

Sum of winning number divided by sum of entry number for each customer in practice scenarios (not real money)



Modifying the Given Data (using mean or sum value)

ID	Seq No.	Col A	Col B
X11	1	2	3.5
X11	2	5	7.3

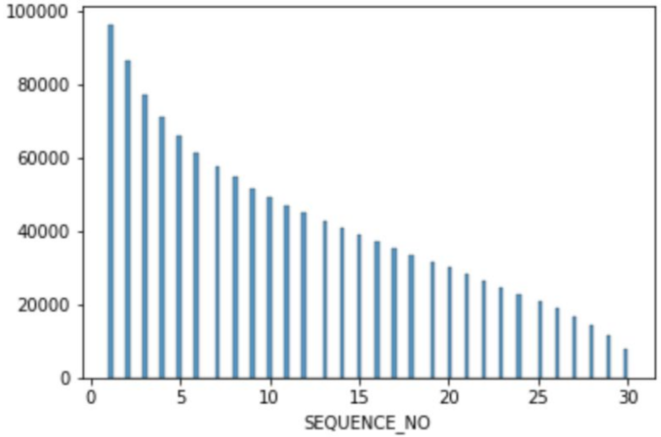
ID	Count	Col A	Col B
X11	2	3.5	5.4

In the given dataset, there were more than 1 column for the same customer. So I decided to group them based on the Unique Customer ID. Grouping was done using mean for certain columns and sum for other columns like Number of Active Cycles, Number of Winnings, Number of Entries etc..

How Long Has The Customer Been With Us?

new feature addition

- Sequence numbers are data points of the same customer taken at different intervals of time
- As the data is collected at equal intervals of time, more the number of data values for a customer, longer the time the customer has been this
- Customers who have been there for a long time tend to be more loyal when compared to new customers
- Higher they stay on the app, higher the chances of them to start playing. But this might also include the people who are not active. So we can add a new variable for the **number of sequence intervals the customer was active based on ACTIVE_YN**



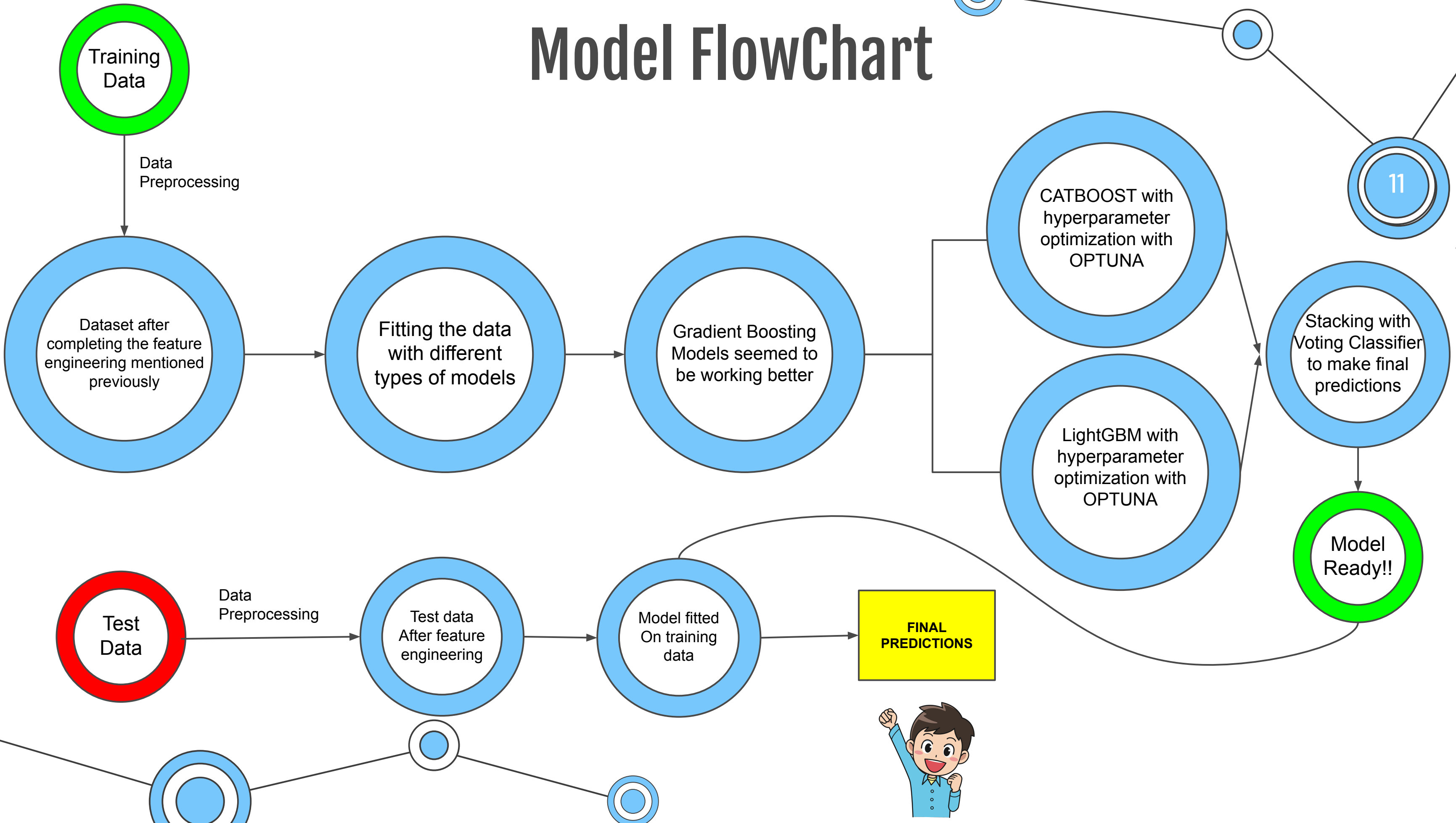
Win Percentage

The customer’s decision of whether to use the app or not depends highly on how much he has won before. If he has been in loss in the past few games, then there is very less chance that he will put more money into the app. His trust will reduce. That is why I added a new variable win_percentage which is calculated as
NO.OF. WINNINGS / NO.OF ENTRIES
For each customer.

Practice Percentage

For customers who are more skeptic about these types of things - like whether to put their money in the app or not. They will try to do some practice games which does not involve actual money. So there is high chance of them putting their money if their winning percentage in practice games are higher. That is why I added the variable practice_percentage which is calculated by
NO.OF. PRACTICE WINNINGS / NO.OF. ENTRIES
For each customer.

Model FlowChart



Training Data

Data Preprocessing

Dataset after completing the feature engineering mentioned previously

Fitting the data with different types of models

Gradient Boosting Models seemed to be working better

CATBOOST with hyperparameter optimization with OPTUNA

LightGBM with hyperparameter optimization with OPTUNA

Stacking with Voting Classifier to make final predictions

Model Ready!!

Test Data

Data Preprocessing

Test data After feature engineering

Model fitted On training data

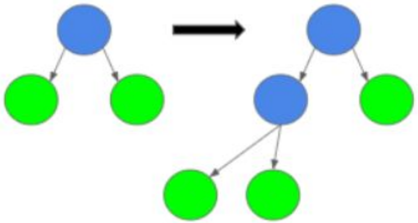
FINAL PREDICTIONS

LightGBM

LightGBM stands for Light Gradient Boosting. Light is due to the high speed execution of the model.

LightGBM model uses **Gradient Based One Side Sampling** to filter out data instances. It downsamples the instances based on their gradient values.

A leaf node with lower gradient i.e less error will not be considered for the next iteration. The leaf node with the higher gradient will be used to make split for the next iteration.

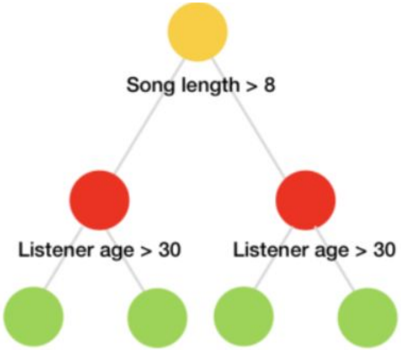


CatBoost

CatBoost stands for Categorical Boosting. The primary benefit of the CatBoost is support for categorical input variables. It also provides an improvement in computation time of the model.

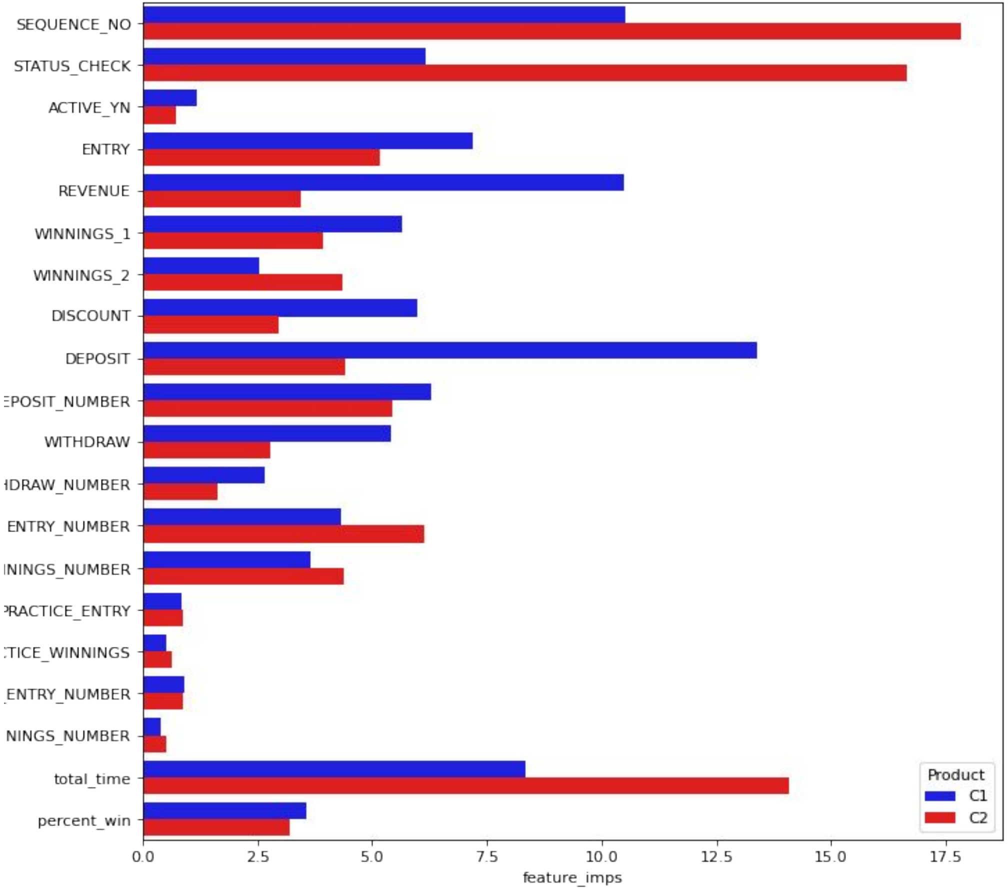
Catboost uses a new technique called **Minimal Variance Sampling** which is similar to weighted sampling of stochastic gradient descent.

Unlike other gradient boosting algorithms, catboost implements symmetrical trees concept. This helps in faster computation.



Feature Importance Plot

As you can see from the plot, SEQUENCE_NO, STATUS_CHECK, REVENUE, DEPOSIT, TOTAL_TIME, ENTRY_NUMBER seem to be having the more importance in the prediction, when compared to other variables in the dataset.



Model Evaluation

Root Mean Squared Error

The model is evaluated based on the Root Mean Squared Error value. The formula for calculating RMSE is given by:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

First, we validated the model in the given training dataset using K-Fold Cross Validation while checking for the best parameters. Then we fitted it on the training data to make the final predictions.

5.04 (Y1)

123.21 (Y2)

Cross Validation Score

68.51

Public Leaderboard

67.50

Private Leaderboard

Future Prospects

Anomaly Detection

The performed anomaly detection using Isolation forest didn't seem to improve the predictions. We could have segregated the values into different segments based on K-Means and added it in a separate column.

Distribution of Variables

Could have looked more into the distribution of variables and check and remove the variables that are highly correlated.

Predictor Variable Relations

Didn't think more about the relation between Y1 and Y2. Could have just used Y1 to predict the value of Y2 or could have used Y1 while creating the model for Y2. This might be a deal breaker.

Choosing between SUM or MEAN

Could have created 2 different models one with the sum of values for each customer and another with mean of values of each customer and then weight average the values. This would have also helped in reducing overfitting if any.

Thank You