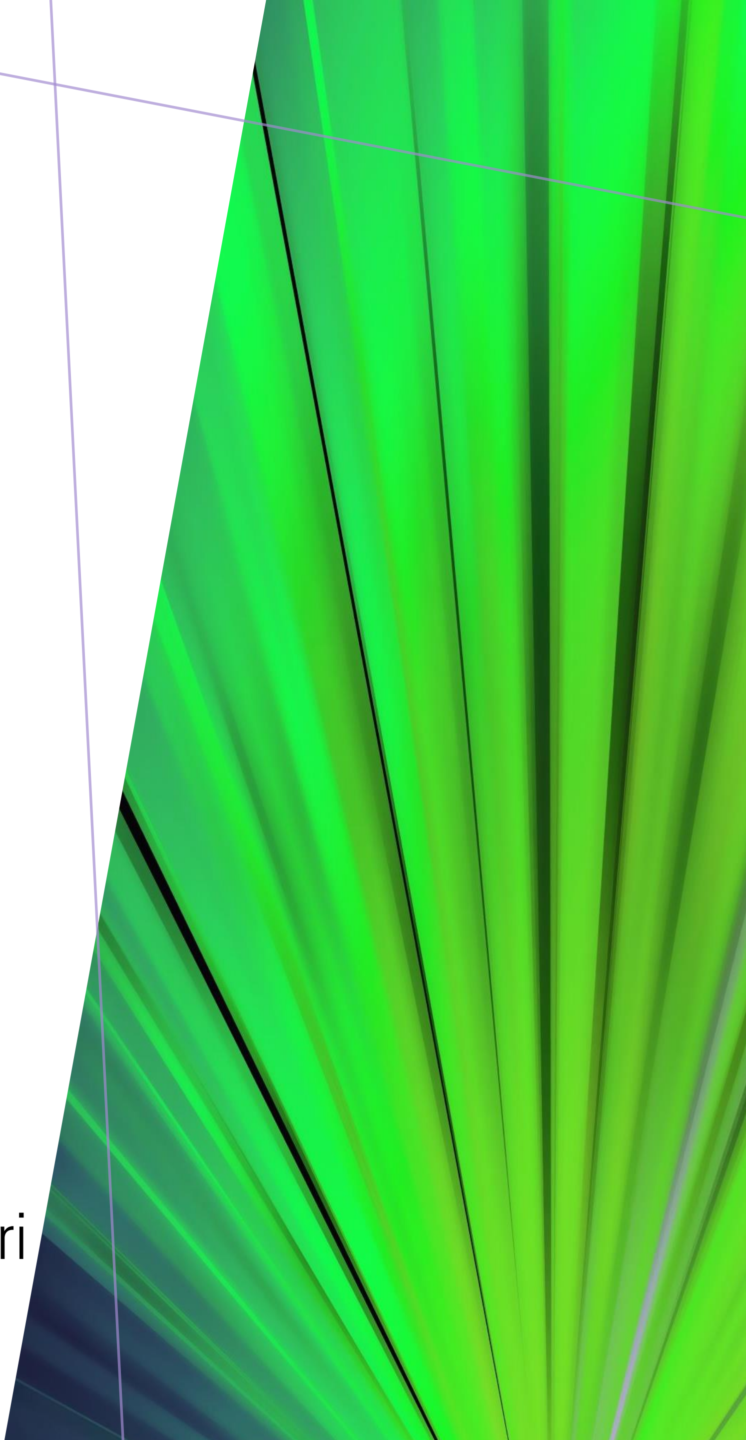
BEYOND ANALYSIS



COR-239
Rajat Maheshwari
Akshat Jain
Madhukar

PROBLEM STATEMENT

- Estimating customer value and extrapolating the existing value into future for the online skill-based gaming
- Data set includes multiple customer behavioral features and their Temporal variation as inputs, and customer values and temporal extrapolation as targets
- We used average mean square error for the targets as our evaluation metric

FEATURE ENGINEERING

- FROM FIGURE 1 WE CAN SEE THE CORRELATION BETWEEN THE PAIR Y2 ,SEQUENCE_NO IS HIGH .

- IN FIGURE 2, WE CAN OBSERVE THAT THE AVERAGE VALUE OF Y2 FOR A CUSTOMER IS INCREASING AS THE SEQUENCE_NO INCREASES.

- BASED ON THE OBSERVATIONS WE HAVE ENGINEERED A NEW FEATURE WHICH STORES THE MAXIMUM SEQUENCE NUMBER FOR A PARTICULAR CUSTOMER

- MULTIPLE TIME STAMPS OF THE SAME CUSTOMER WAS GIVEN IN THE DATASET, SO WE TOOK THE MEAN OVER THE SAMPLE WHICH BELONG TO SAME CUSTOMER

Figure 1

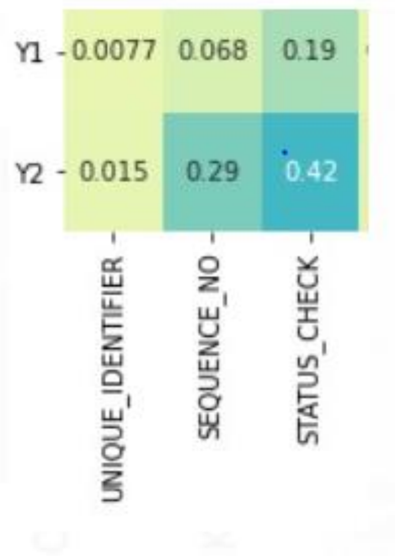
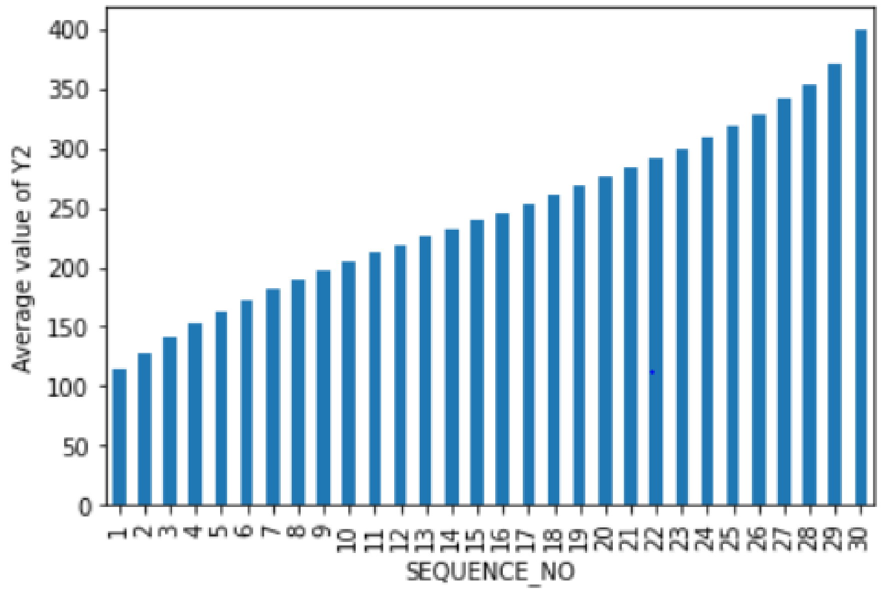


Figure 2

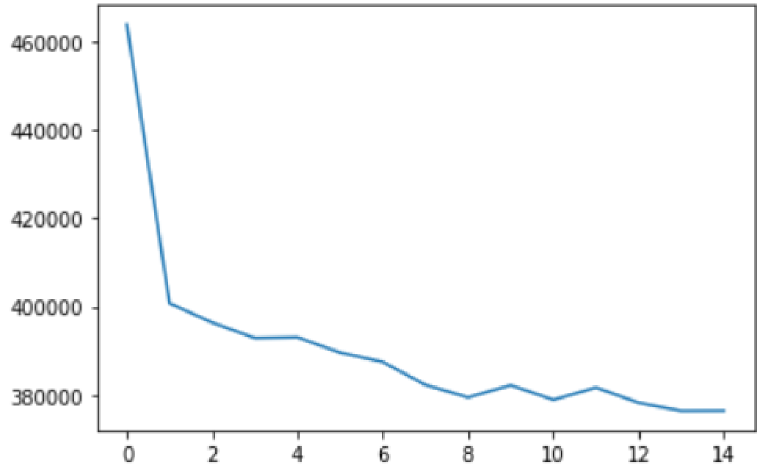


METHODS EXPLORED

- Deep Learning
  - We have trained two models out of which model1 had y1 as target and model2 had both y1 and y2 as their targets.
  - When generating the prediction for test we considered prediction for y1 from model1 and prediction for y2 from model2.
  - The reason for doing this , we observed higher mse when model for predicting Y2 was trained alone , but observed lower mse when both y1 and y2 were trained together.

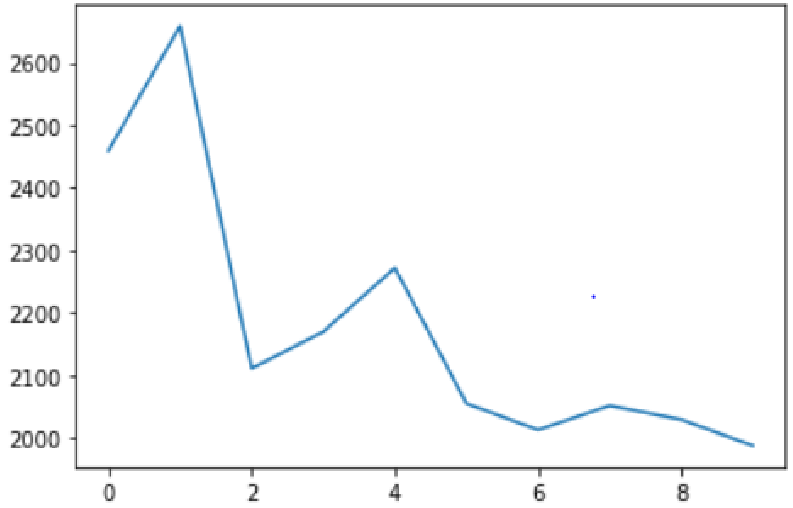
MSE vs Epochs GRAPHS

```
Iteration (epoch) 13 / 15 , Loss 7039.81 Train mse 556634.41 Test mse 196147.18
Iteration (epoch) 14 / 15 , Loss 8357.85 Train mse 556693.67 Test mse 196148.84
```



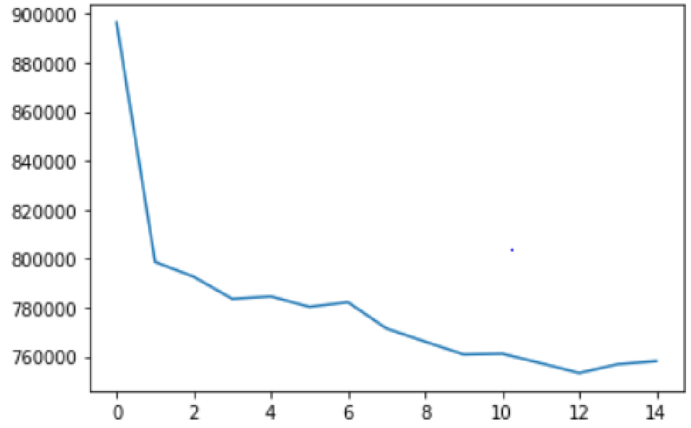
When Y1 and Y2 was used as Targets

```
Iteration (epoch) 8 / 10 , Loss 42.81 Train mse 3099.09 Test mse 959.48
Iteration (epoch) 9 / 10 , Loss 13.73 Train mse 3025.65 Test mse 949.34
```

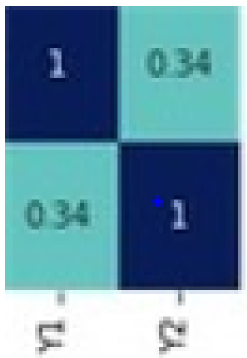


When Y1 was used as Target

```
Iteration (epoch) 13 / 15 , Loss 15905.72 Train mse 1119102.27 Test mse 394473.67
Iteration (epoch) 14 / 15 , Loss 14743.59 Train mse 1122123.21 Test mse 394004.74
```



When Y2 was used as Target



METHODS EXPLORED

XGBoost Regressor	Random Forest Regressor
• We tried tuning hyperparameters	• We tried tuning hyperparameters

Among these three methods we observed that Deep Learning gave us the lower overall MSE

TOOLS USED

- For model building we used:
  - PyTorch
  - Scikit-Learn
  - XGBoost library
- For visualization:
  - Seaborn
  - Matplotlib
- For data preparation:
  - Pandas
  - NumPy
- For performance measurement:
  - MSE for Deep Learning model
  - RMSE for Machine Learning model

RESULTS

- We estimated customer values and temporal extrapolation which are continous in nature
- When Y1 was trained alone approximate MSE value was **3025**
- When Y2 was trained alone approximate MSE value was **1122123**
- When both Y1 and Y2 was trained simultaneously approximate MSE value was **556693**
- Due to this we considered Y2 from the combined model and Y1 from the stand-alone model for prediction of the test data
- Using the above techniques, we achieved an RMSE of **67.85363**

CHALLENGES

- Multiple time stamps of the same customer was given in the dataset
- After averaging the data over the customer ID there was a dramatic reduction in the dataset
- Due to this our testing RMSE increased

LESSONS LEARNED

- We learned how to handle Multiple time stamp data
- We used combined results from two different models to do the prediction on the test data, these kind of approaches can be used to increase performance of the model

FUTURE PROSPECTS

- By performing extensive EDA, we can remove columns which have less impact
- We observed there more outliers in the features Winnings_1 , upon removing may improve the performance of the model.
- Since the target variables involve temporal extrapolation, we can build DL models which can handle time series data effectively like RNN and LSTM based models.