






# IST 557 Project Progress-I

Mihir Mehta (mzm6664) & Mudit Garg (mxg5783)

## Best performance:

The best performance- 0.98537

The corresponding snapshot is:

|   |           |  |         |    |     |
|---|-----------|--|---------|----|-----|
| 1868  | mihirpsu  |  | 0.98537 | 2  | 4m  |
| 1869  | IST557_A3 |  | 0.98537 | 4  | now |
| <b>Your Best Entry</b>         |           |  |         |    |     |
| You advanced 136 places on the leaderboard!   |           |  |         |    |     |
| Your submission scored 0.98537, which is an improvement of your previous score of 1.01230. Great job!           |           |  |         |    |     |
|  <a href="#">Tweet this!</a> |           |  |         |    |     |
| 1870  | hedemann  |  | 0.98584 | 47 | 5mo |

## The Best method:

XGBoost Regressor:

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

## Motivation:

<https://www.kaggle.com/dlarionov/feature-engineering-xgboost>

With parameters

```
max_depth=8,  
n_estimators=1000,  
min_child_weight=300,  
colsample_bytree=0.8,  
subsample=0.8,  
eta=0.3
```

Other methods tried:

- 1) Time Series Approach: We understood that it is decreasing time series with seasonality. We need to fine tune ARIMA model with (p, q, r). However, given number of time series we need to predict, we could not scale the same.
- 2) Decision Tree: We tried decision tree. It was giving unstable results as expected.
- 3) Random Forest: We tried random forest which gave very bad results i.e. 2.85. But then we did not do much feature engineering for the same.

**Summary and plans:**

- 1) From using <https://www.kaggle.com/dlarionov/feature-engineering-xgboost> as a motivation, we did majority of similar feature engineering.
- 2) We found that adding features help to cluster similar products together. Also, it helps to capture decreasing trend and seasonality
- 3) Going forward, we would like to see if we can add additional features to improve accuracy
- 4) We will work towards building ensemble model using different XGBoost regressors
- 5) We will explore if LSTM network can be used to get better results
- 6) We will also see if we can improve external data sources to get better results.
- 7) Additionally, we would like to explore test dataset further for more insights and use test dataset inspired features in our model.

**Contributions:**

Mudit Garg: 50%

Mihir Mehta: 50%

**References:**

- 1) <https://www.kaggle.com/kyakovlev/1st-place-solution-part-1-hands-on-data>
- 2) <https://www.kaggle.com/minhtriet/a-beginner-guide-for-sale-data-prediction>
- 3) <https://www.kaggle.com/dlarionov/feature-engineering-xgboost>