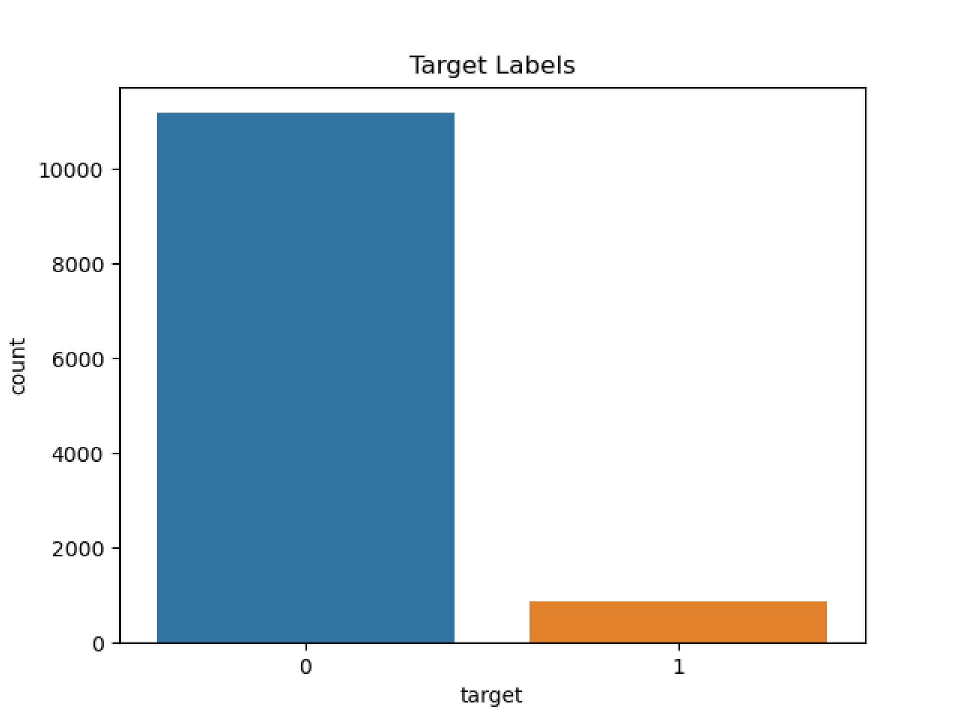
Receipt Matching Data Science Challenge

- Sai Kasyap Kamaraju

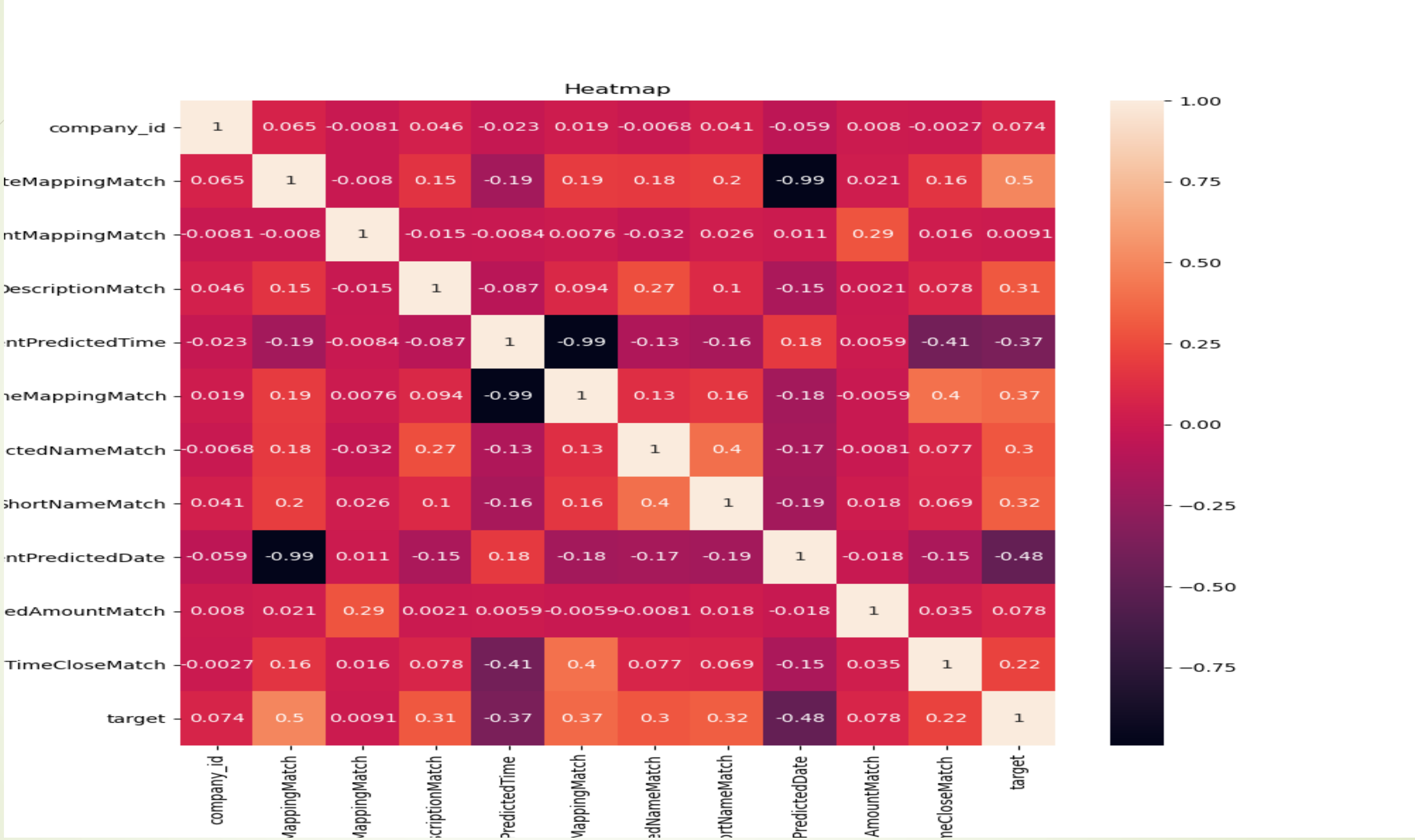
(Sr. Data Scientist)

Goals & Dataset Description

- Automatically match the receipt images with the transactions associated.
- In the app when the customer takes a picture of a receipt, the app provides a list of transactions likely to match the receipt, goal is map it to the correct transactions from the list.
- Highly imbalanced dataset, with only 857 correct matches and 11177 incorrect matches.
- Correct matches are defined if matched_transaction_id equals feature_transaction_id.

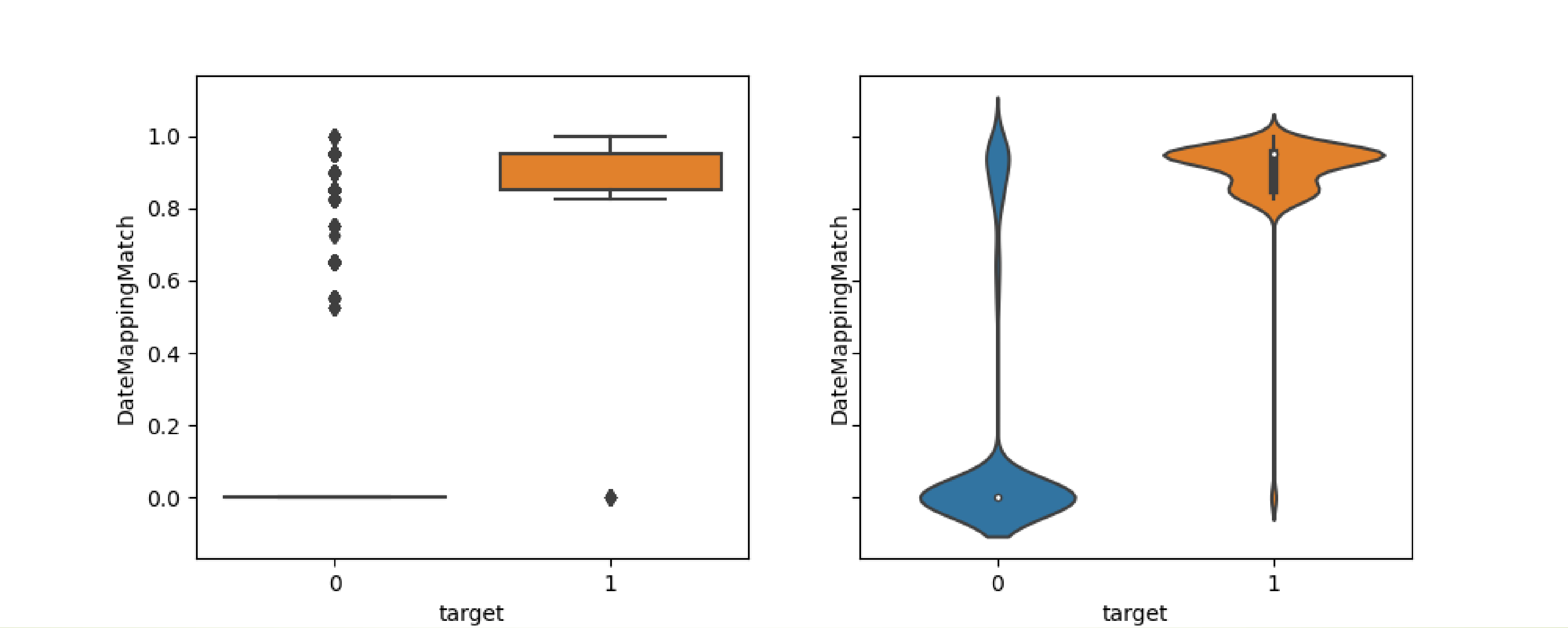


EDA



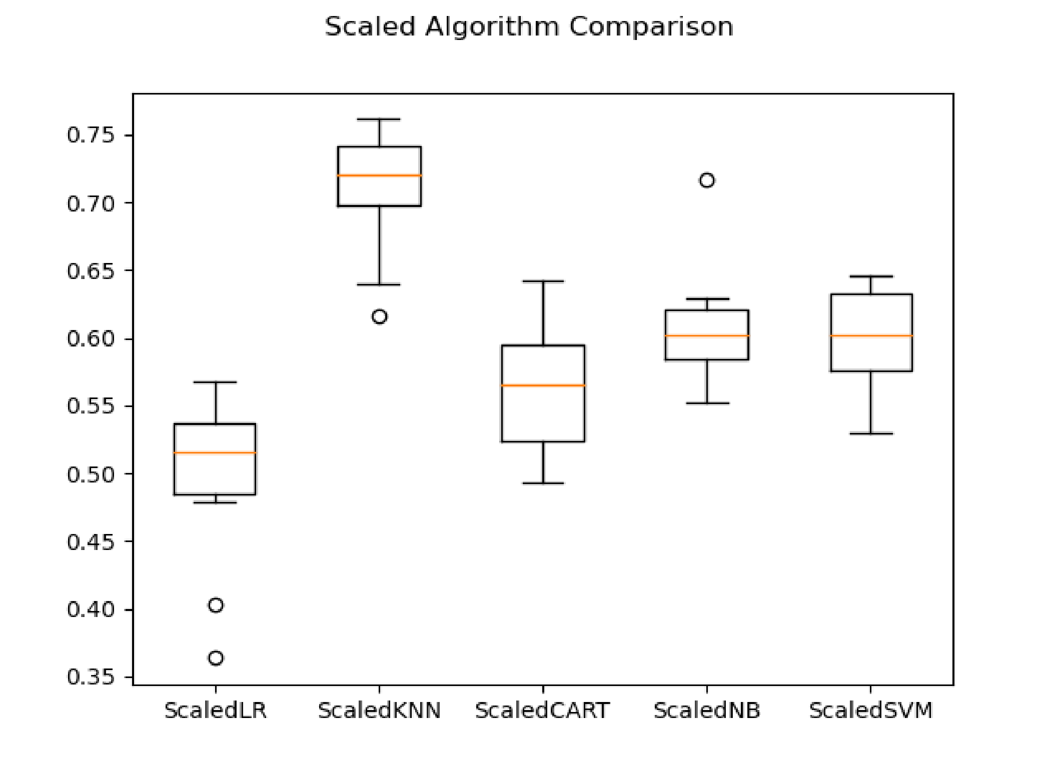
DifferentPredictedTime, DifferentPredictedDate can be dropped as they are negatively correlated with respect to target.

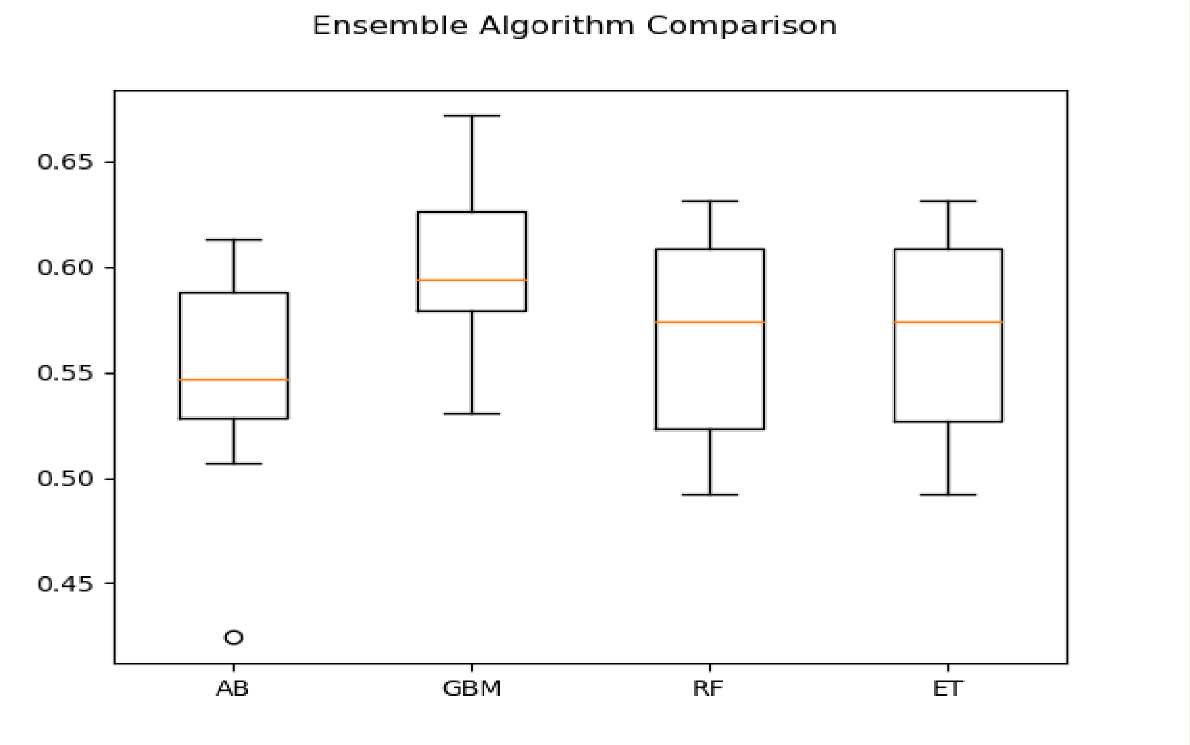
EDA – Contd..



Boxplot & volinplot for DateMappingMatch with respect to target.

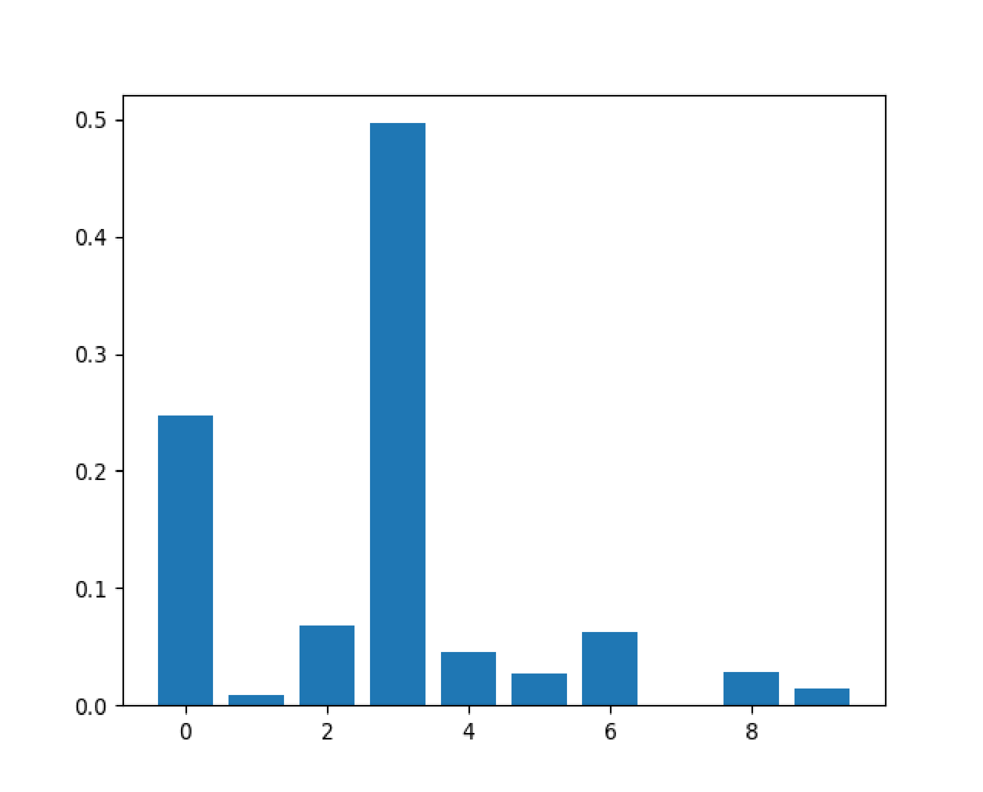
Comparison Of Different Algorithms

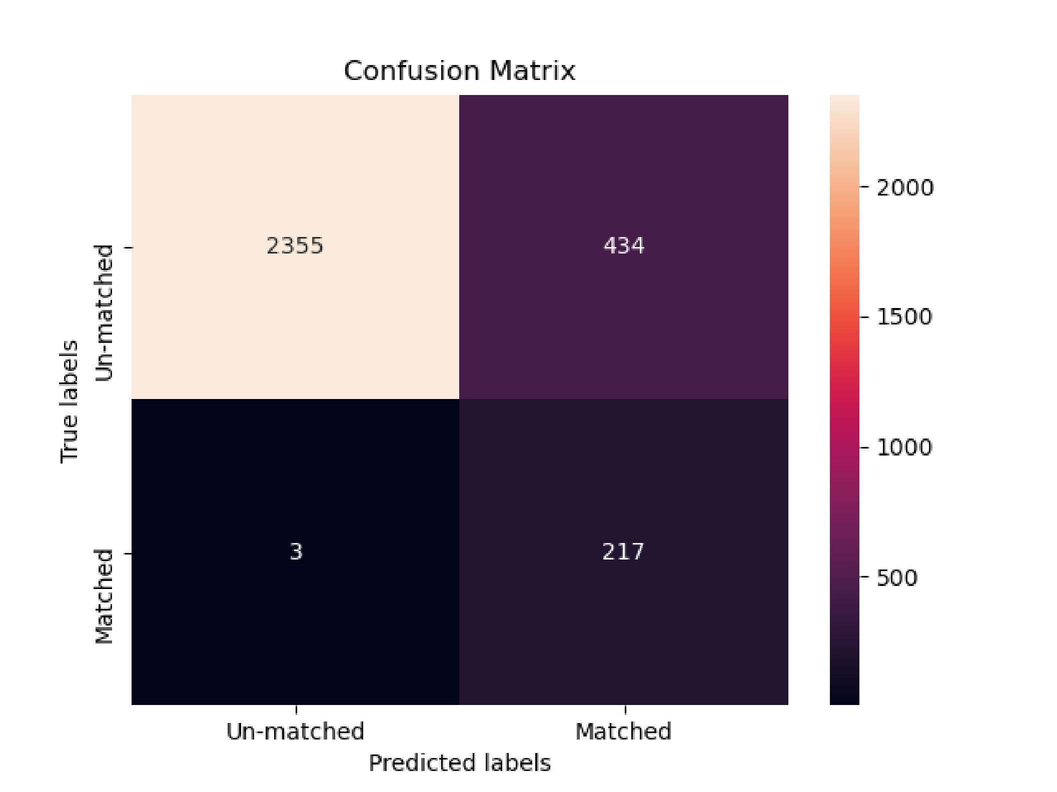

Scaled Algorithm Comparison


Ensemble Algorithm Comparison

KNN performs better than the others…

Use of XGBOOST





Feature importance using Xboost Classifier.

Using XBOOST gives 0.99 recall.

- **Future steps:**
- Hyperparameter optimization of XGBOOST, to improve the metrics further.
- Use of Oversampling/ undersampling technniques to balance the dataset.
- Use of deep learning based approaches for larger dataset.
- **Conclusion:**
- Recall is chosen as the main evaluation metric to consider.
- Hyper-Parameter optimization for Xgboost and others might not significantly improve the results as the dataset is imbalanced.
- Dropping negative correlated features/ outliers didnt produce better results.
- Data transformation might be needed after understanding of each of the column features.
- Dimensionality reduction / DL based approaches can be explored.

Thank You!