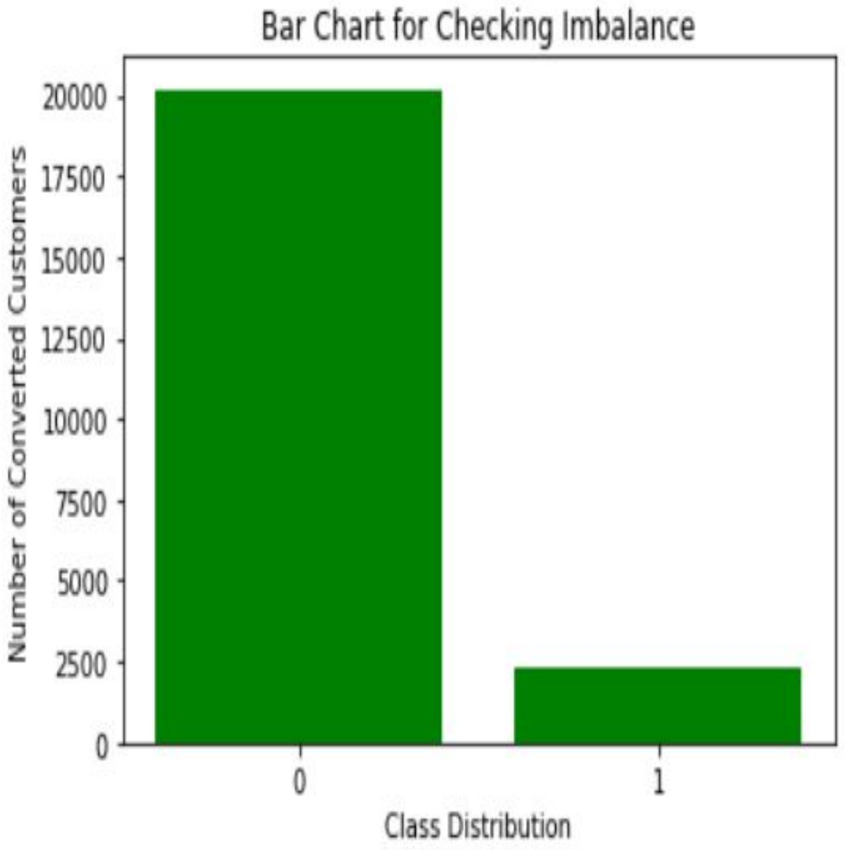
Using Data to bring Customers Home

Alefiya Naseem

Devanshi Gariba

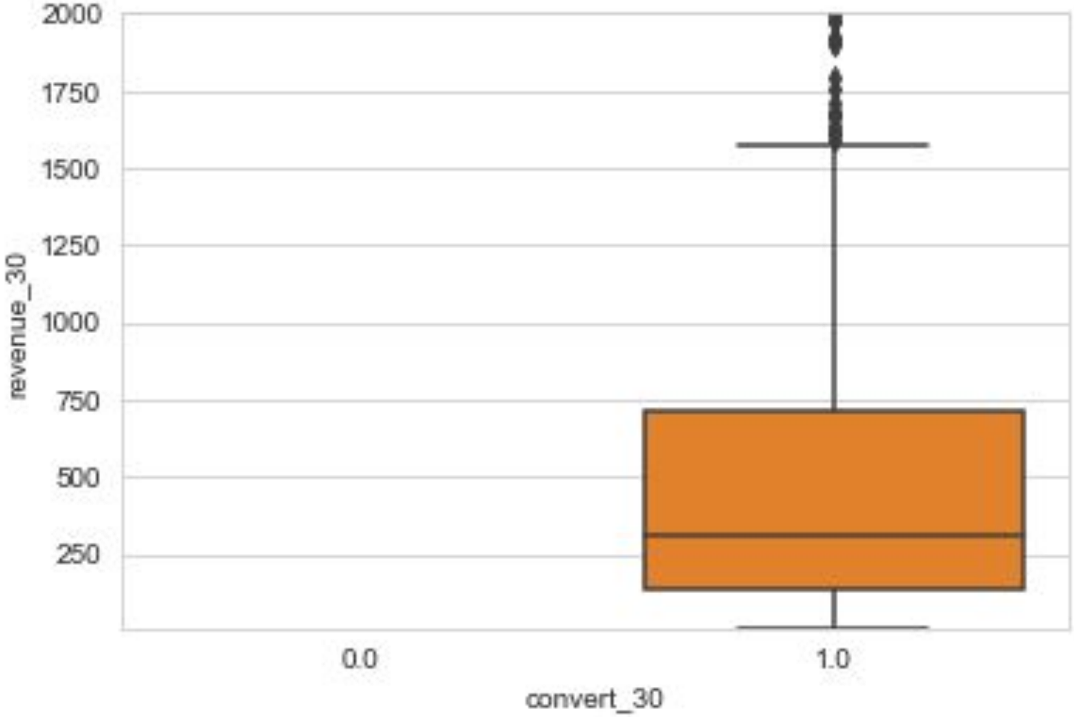
Target variable distribution and Statistics

- Mean monthly revenue for customers who converted is $721
- Distribution of revenue- 75% of the customers spend less than $750. Min - Max revenue range is $5 - $35000
- High Class imbalance – only 10% of all customers will actually purchase in the next month

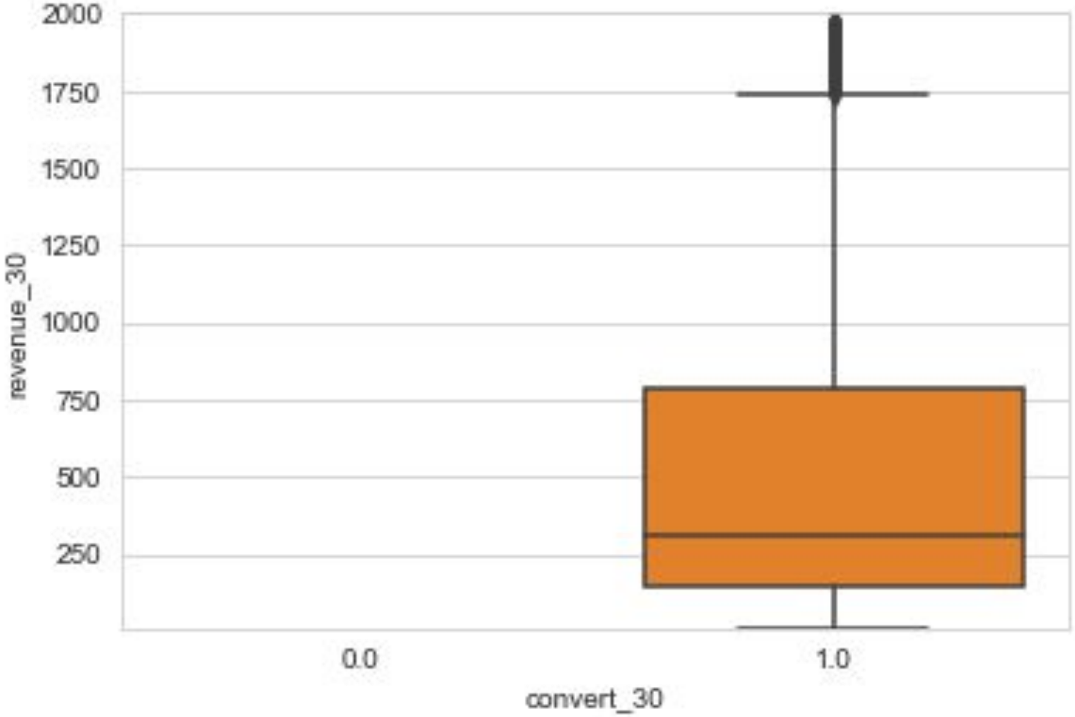


Test-Train Distribution: Covariance Shift Check

Test Revenue Distribution



Train Revenue Distribution



Approaches

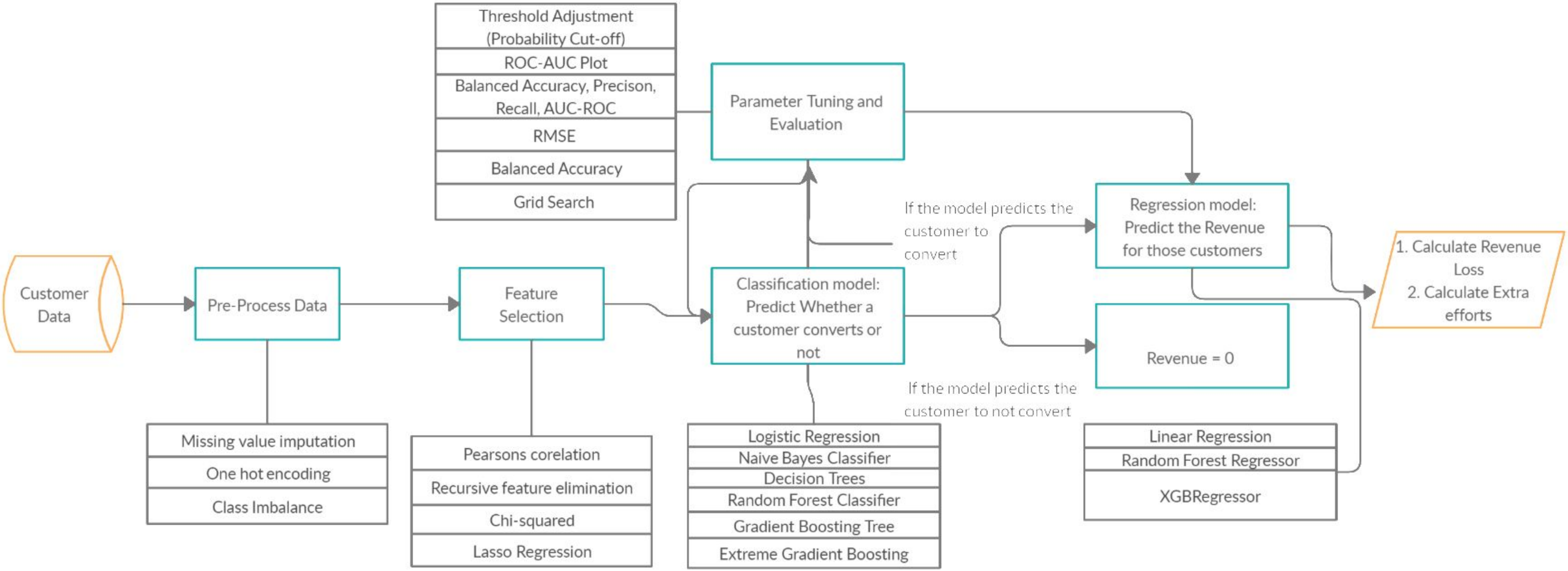
1. Independent classification and regression models to predict revenue generated and whether a customer converted or not.

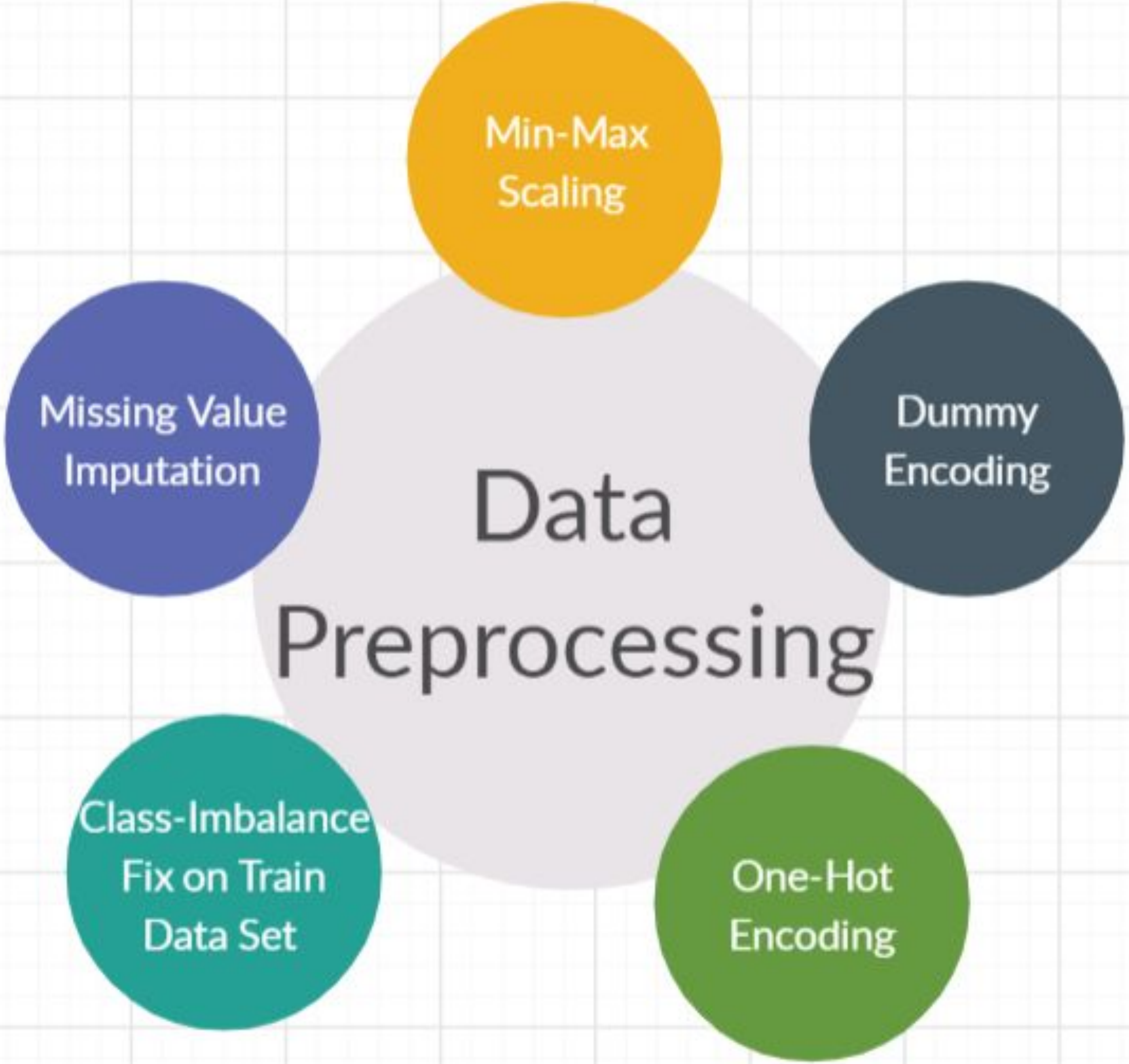
   Revenue_RMSE = $ 365

   Balanced_Accuracy = 67%

2. Channelise the output of the classification model to the regression model and only predict the revenue for customers who actually make a purchase.

Model Pipeline


Customer Data
Pre-Process Data
Feature Selection
Classification model: Predict Whether a customer converts or not
Parameter Tuning and Evaluation
Threshold Adjustment (Probability Cut-off)
ROC-AUC Plot
Balanced Accuracy, Precison, Recall, AUC-ROC
RMSE
Balanced Accuracy
Grid Search
If the model predicts the customer to convert
If the model predicts the customer to not convert
Regression model: Predict the Revenue for those customers
Revenue = 0
1. Calculate Revenue Loss
2. Calculate Extra efforts
Missing value imputation
One hot encoding
Class Imbalance
Pearsons corelation
Recursive feature elimination
Chi-squared
Lasso Regression
Logistic Regression
Naive Bayes Classifier
Decision Trees
Random Forest Classifier
Gradient Boosting Tree
Extreme Gradient Boosting
Linear Regression
Random Forest Regressor
XGBRegressor



Data Preprocessing

Min-Max Scaling

Dummy Encoding

One-Hot Encoding

Class-Imbalance Fix on Train Data Set

Missing Value Imputation

Feature Selection Techniques

Pearson’s Correlation

Lasso Regression
Random Forest Regressor
XGboost

Recursive Feature Elimination

Chi-Square Test

Insights on Feature Sets

Active and In progress category customers are much more likely to purchase then Enrolled category customers

Number of Search Terms:
Low v/s High search terms. High number suggests non-specific and general search terms, low suggests specific terms

Number of ATC's and ATF's:
Customers are more likely to buy if they've already added specific items

Number of Visits:
User segmentation - New users v/s Old users and it can be used in determining and designing new strategies around new and old users. Customers who have visited the site in last 1 month are more likely to purchase

Percentage of Dirty Orders:
Customers with lesser dirty orders are likely to purchase again. Better services can also lead to likely conversion.

Revenue generation increases with quote prices, total revenue in last year, price of items added to cart, sum revenue in past 1 month

Candidate Classification Models

1) Logistic Regression

2) Decision Trees

3) Random Forest

4) Gradient Boosting

5) Naive Bayes

6) XGB Classifier

Parameter Tuning

1) Grid-Search

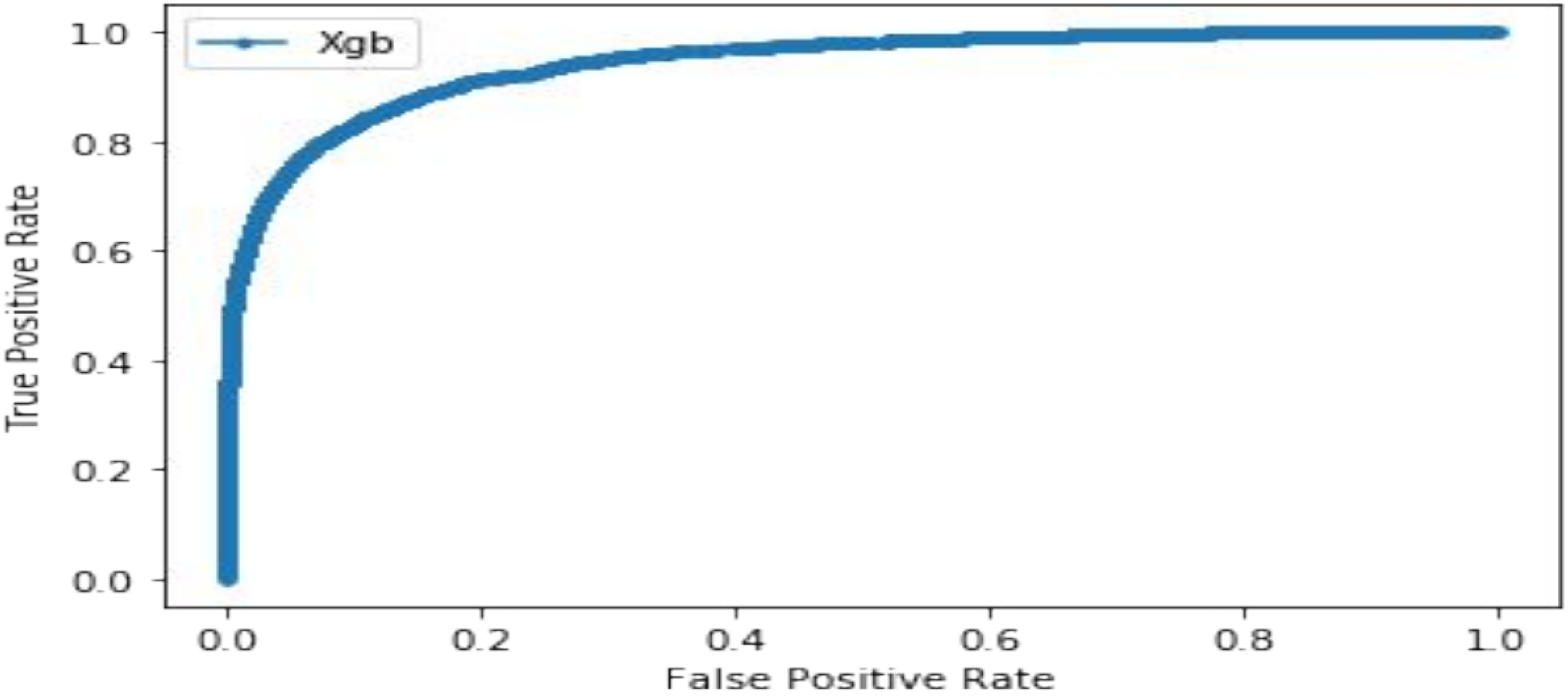
2) Plotted the ROC-AUC curve for Different classification models.

3) Adjusted the threshold to maximize TPR , minimize FPR, maximize balanced accuracy.

Classification Models Comparison

Classifier	Accuracy	F1 Score	Recall	Precision	TNR	Balanced Accuracy
LogisticRegression	75.31	0.79	0.75	0.87	0.24	70.01
DecisionTreeClassifier	76.54	0.80	0.76	0.87	0.26	70.34
RandomForestClassifier	89.05	0.86	0.89	0.85	0.47	0.55
GradientBoostingClassifier	88.58	0.86	0.88	0.85	0.41	0.56
Naive Bayes	86.42	86.45	86.42	86.49	0.37	65.01
XGBClassifier	82.00	0.82	0.79	0.87	0.28	70.27

Best Classifier Model: Extreme Gradient Boosting



Candidate Regression Models

1) Linear Regression
2) XGboost Trees
3) Random Forest

Regression Models Comparison

Regressor	RMSE
Linear Regression	1103
Random Forest Regressor	909
XGBRegressor	507

Motivation for Business Metrics

1. Is RMSE sufficient to understand how the model is really performing?

2. Does it give us an accurate idea of the customers that should be chased v/s the customers that are being chased?

3. Is the model able to capture the revenue lost by not predicting revenue for customers who actually converted?

4. Is the model able to quantify the efforts that wayfair teams will be spending on customers who are actually not converting but the model predicts them to.

Business Metrics for Final Model Evaluation

True_convert	Predicted_convert	True_revenue	Predicted_revenue
1	0	3	0
1	1	39	44
1	0	40	0
1	1	60	82
0	1	0	30
0	1	0	40
0	1	0	50
1	0	20	0
Business Loss		63	
Extra Efforts			120

Business Metrics for Final Model Evaluation

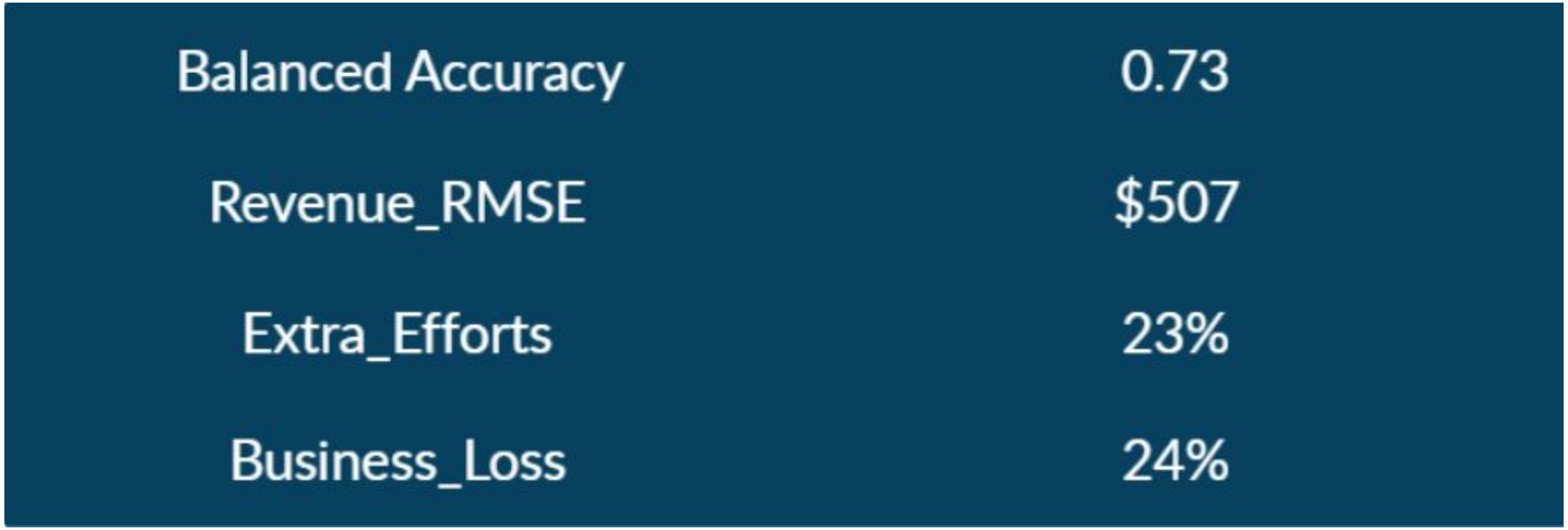
1. Business Loss:

   It is the total revenue that the model failed to capture because classifier input to the final regressor model was 0. Hence revenue predicted for those customers is 0.

2. Extra Efforts:

   The total revenue that is falsely predicted for the customers who did not convert.

Final Results on Test-Data Set



Balanced Accuracy	0.73
Revenue_RMSE	$507
Extra_Efforts	23%
Business_Loss	24%

Web-App Deployment

1. Unlabeled holdout data hosted
2. Predict conversion and revenue for these customers
3. Provides keypoints for improvement for non-converted customers

DEMO:

http://localhost:8080/

2adpro
Happy
HOLIDAYS