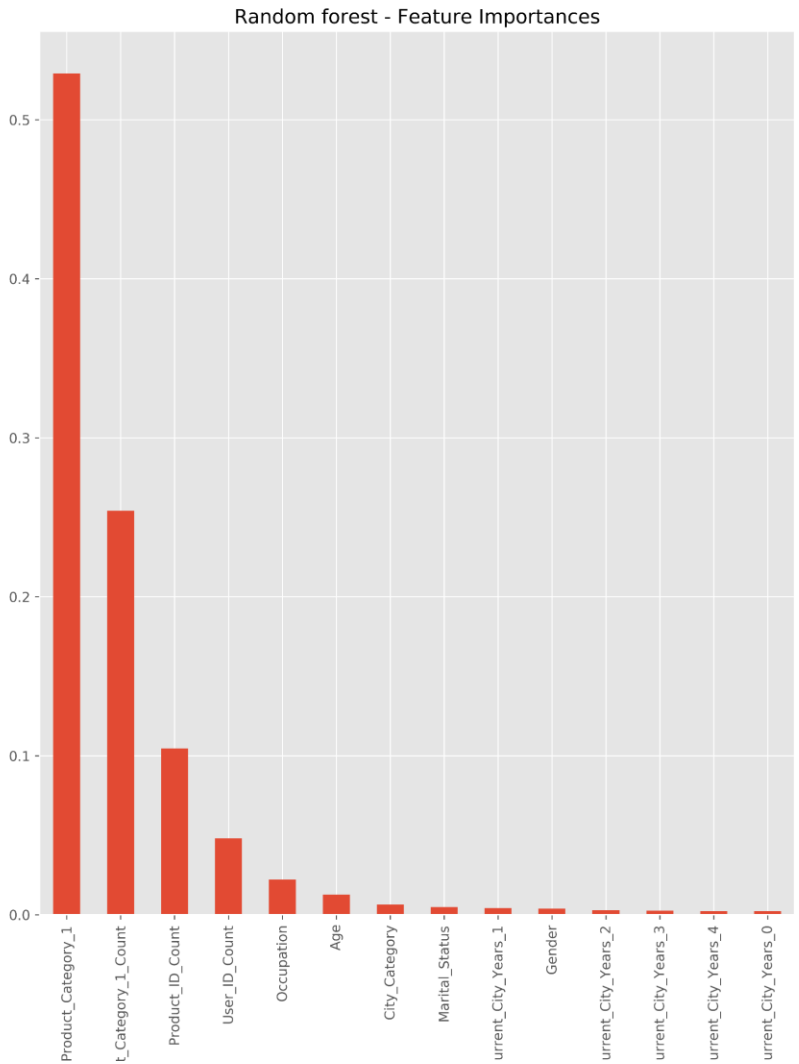
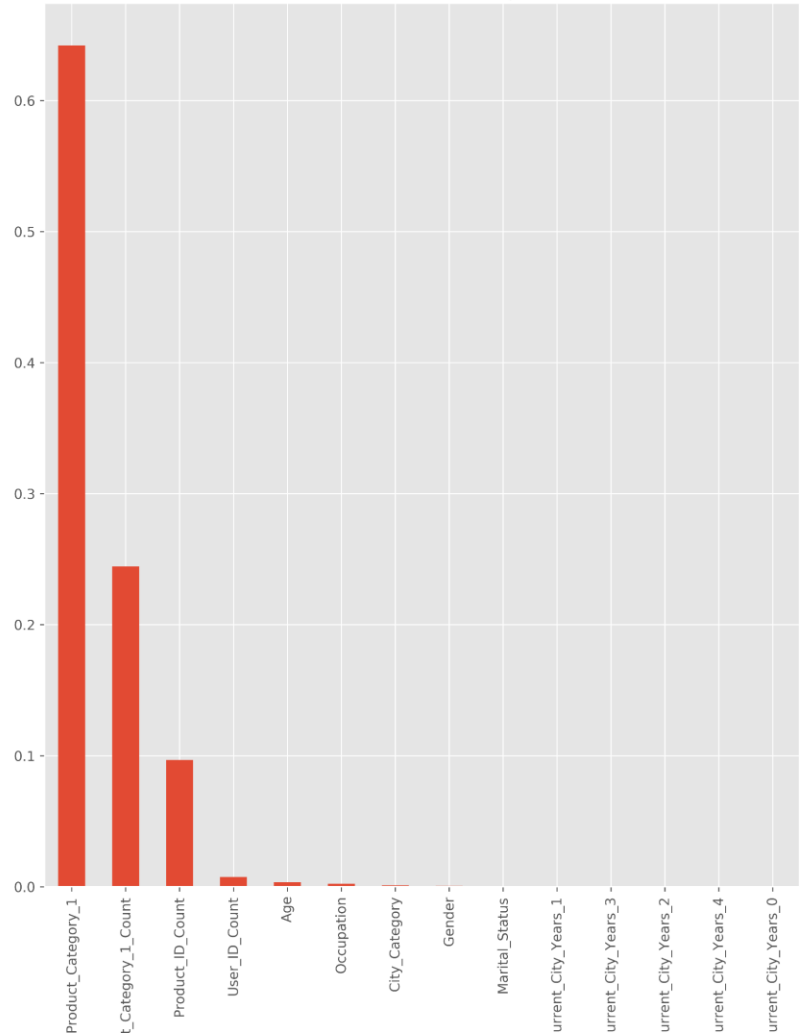
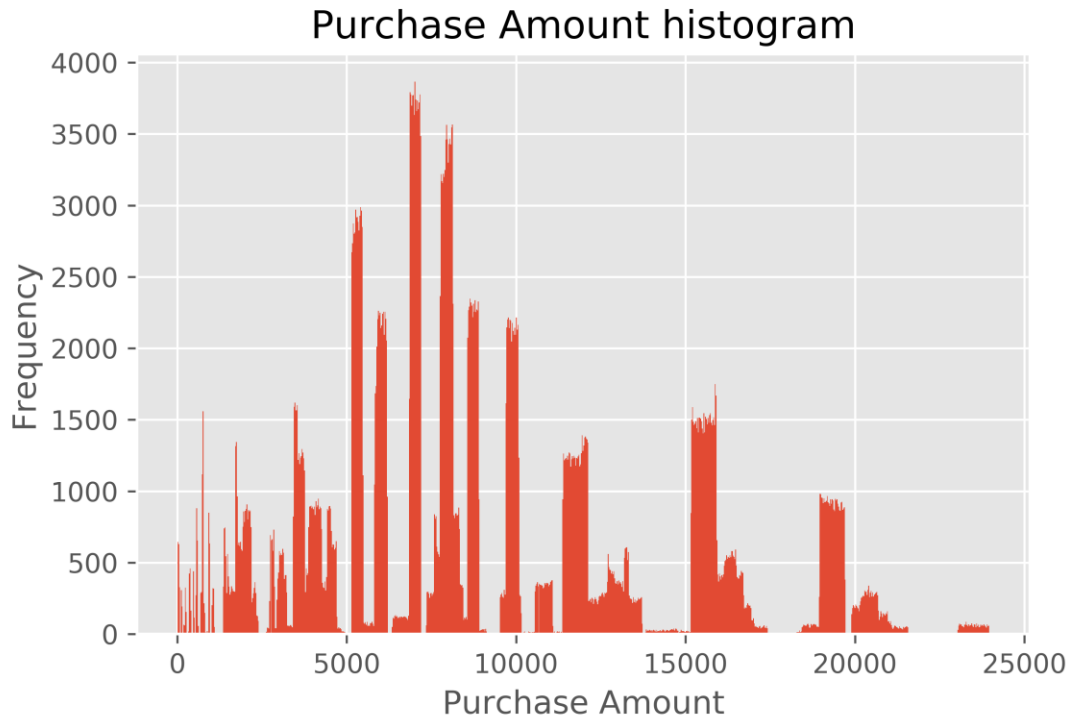
# Statistical Machine Learning – Fall 2019
# Purchase Capacity Prediction based on User demographics and Product information

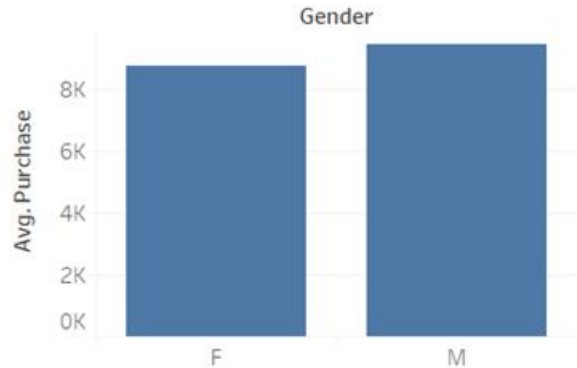
**Keerthiraj Nagaraj**
**Electrical and Computer Engineering, University of Florida**


Random forest - Feature Importances
0.5
0.4
0.3
0.2
0.1
0.0
Product_Category_1
t_Category_1_Count
Product_ID_Count
User_ID_Count
Occupation
Age
City_Category
Marital_Status
urrent_City_Years_1
Gender
urrent_City_Years_2
urrent_City_Years_3
urrent_City_Years_4
urrent_City_Years_0

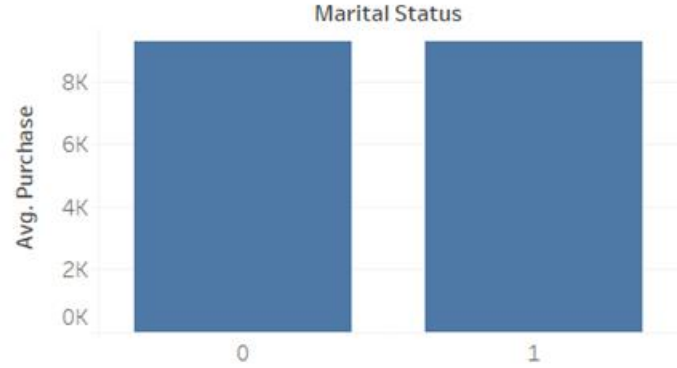

Decision Trees - Feature Importances
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Product_Category_1
t_Category_1_Count
Product_ID_Count
User_ID_Count
Age
Occupation
City_Category
Gender
Marital_Status
urrent_City_Years_1
urrent_City_Years_3
urrent_City_Years_2
urrent_City_Years_4
urrent_City_Years_0


Purchase Amount histogram
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
0
5000
10000
15000
20000
25000
Purchase Amount

## Average Purchase Vs Gender

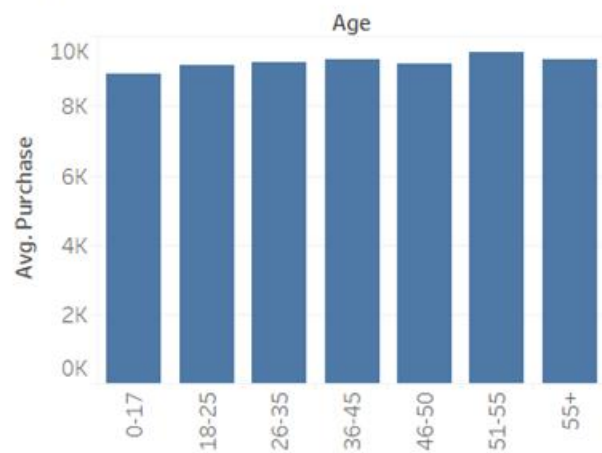



## Average Purchase Vs Marital Status

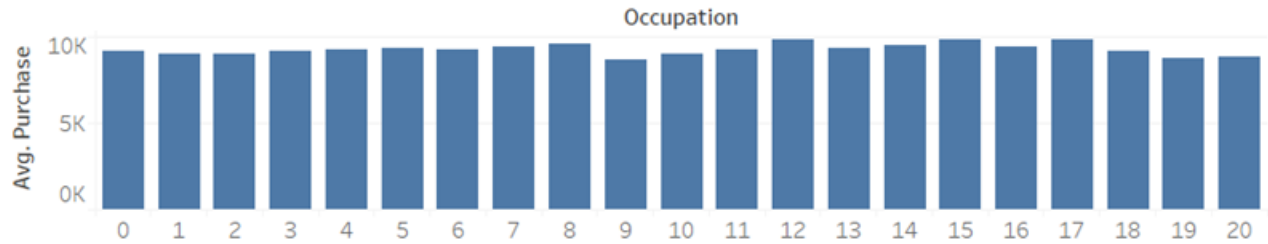



## Average Purchase Vs City Category

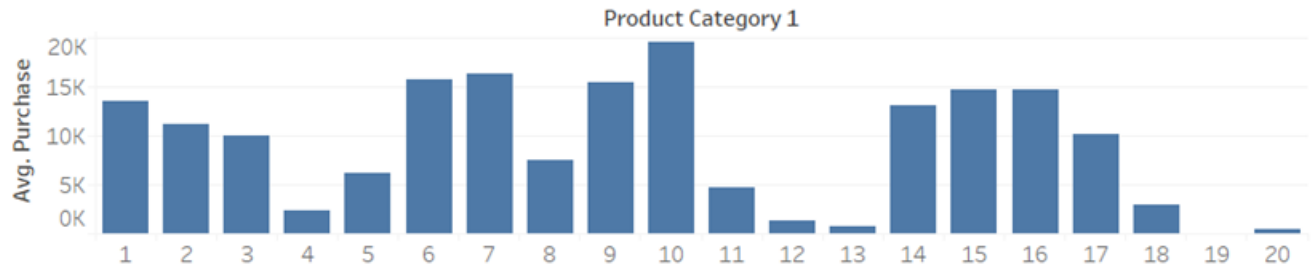

## Average Purchase Vs Age



## Average Purchase Vs Product Category



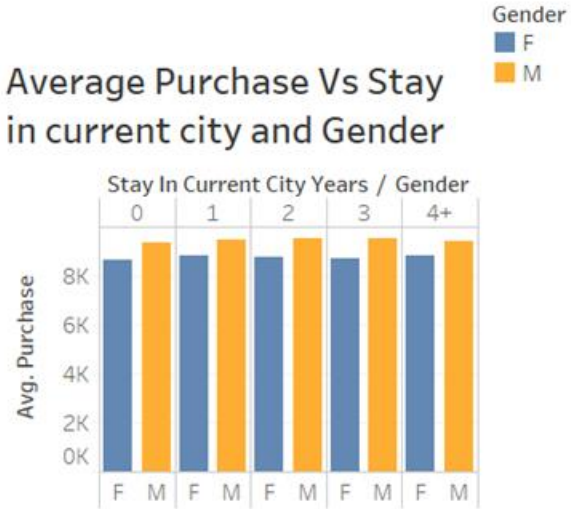

## Average Purchase Vs Product Category



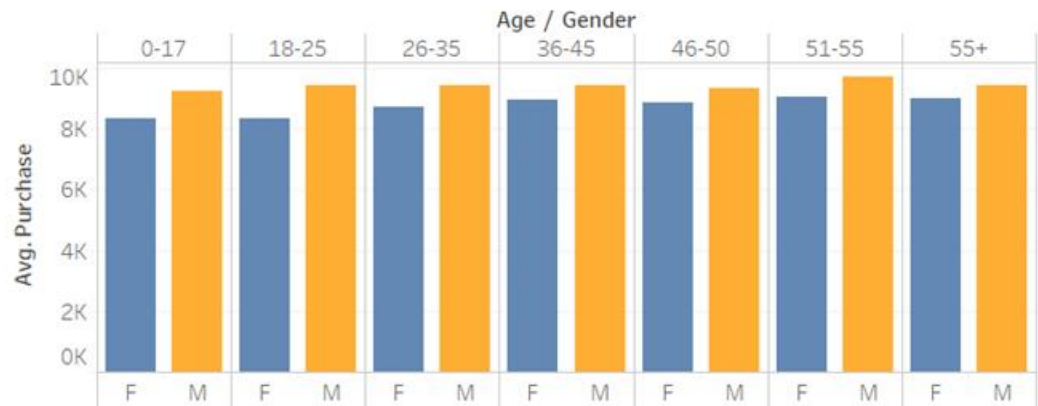

## Average Purchase Vs Product ID (with >1000 data entries)

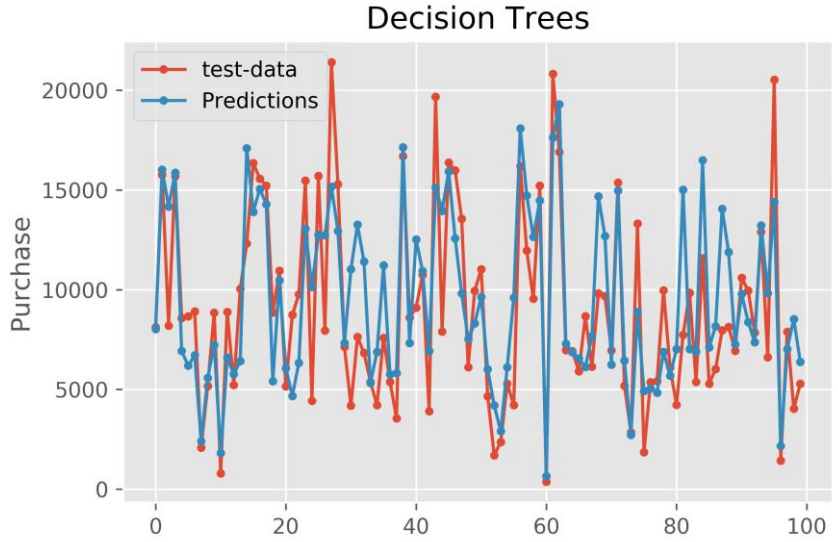
## Average Purchase Vs City Category & Gender & Marital status

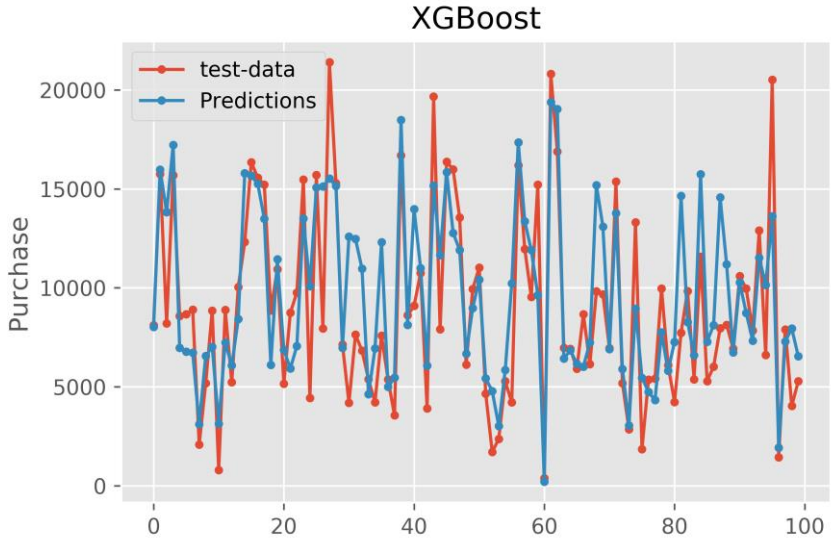

## Average Purchase Vs Stay in current city and Gender

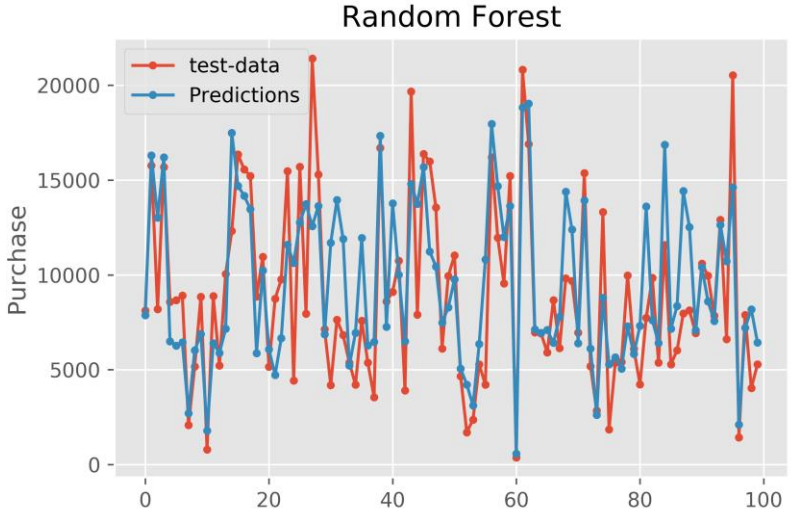


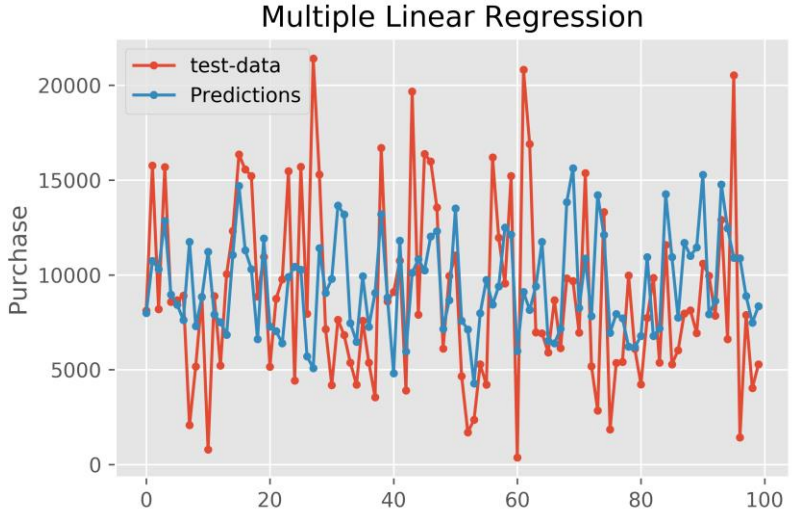

## Average Purchase Vs Age groups and Gender

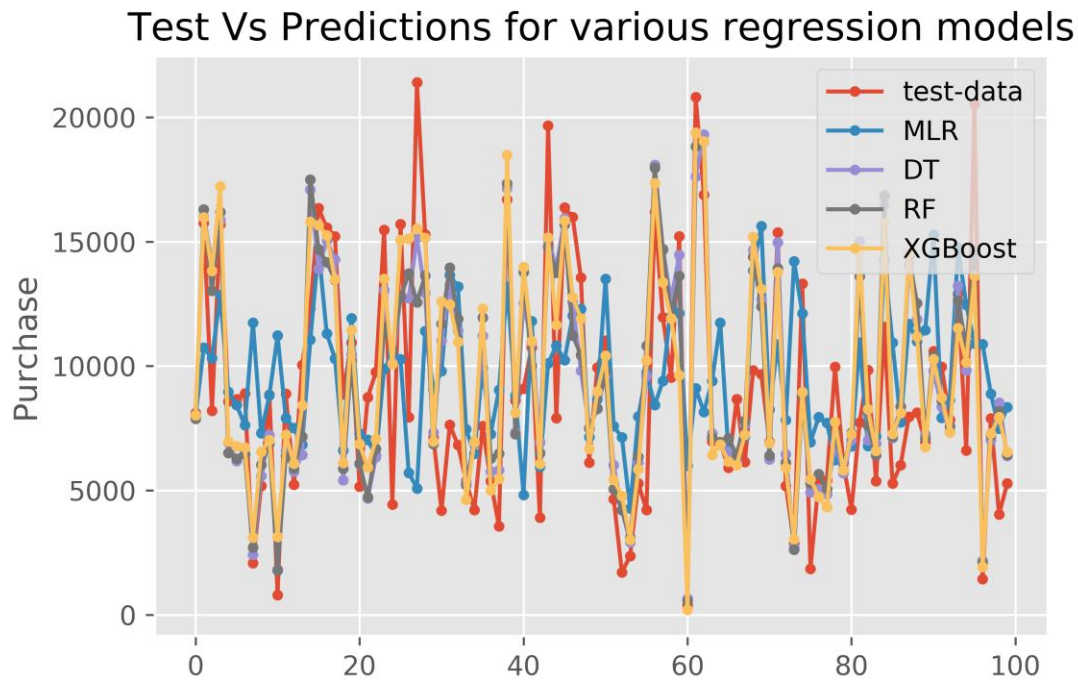


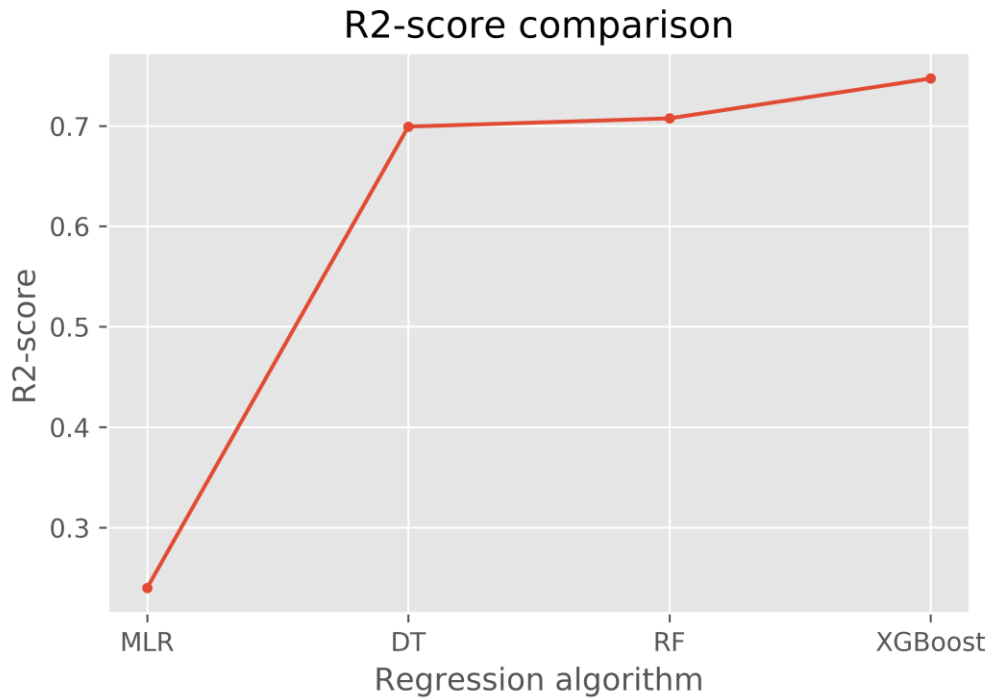

Decision Trees
test-data
Predictions
Purchase
20000
15000
10000
5000
0
0
20
40
60
80
100

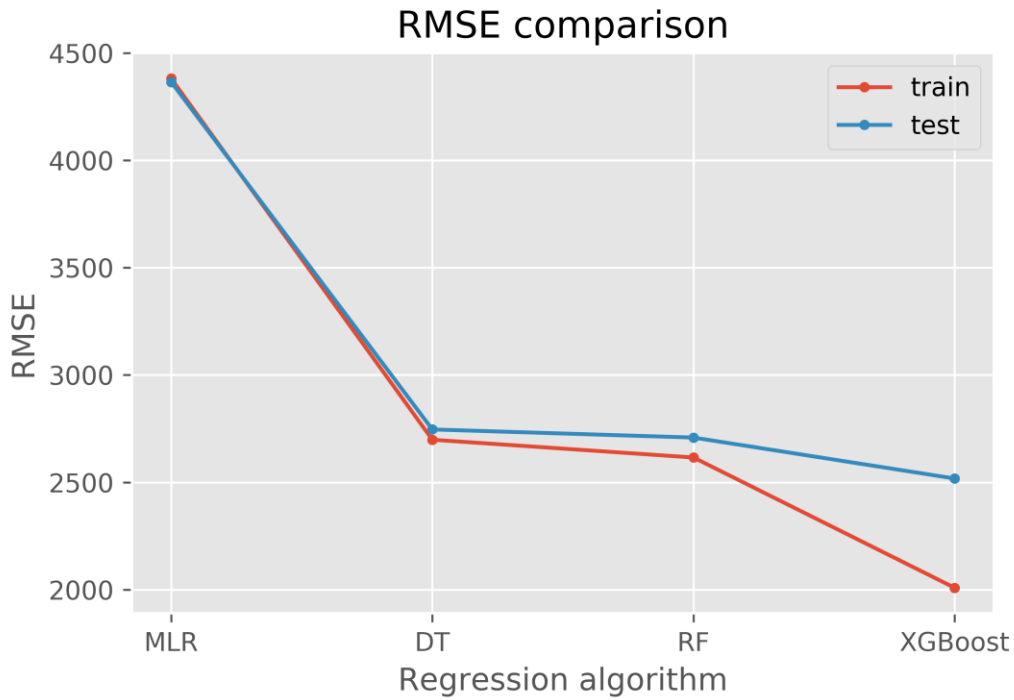

XGBoost
test-data
Predictions
Purchase
20000
15000
10000
5000
0
0
20
40
60
80
100

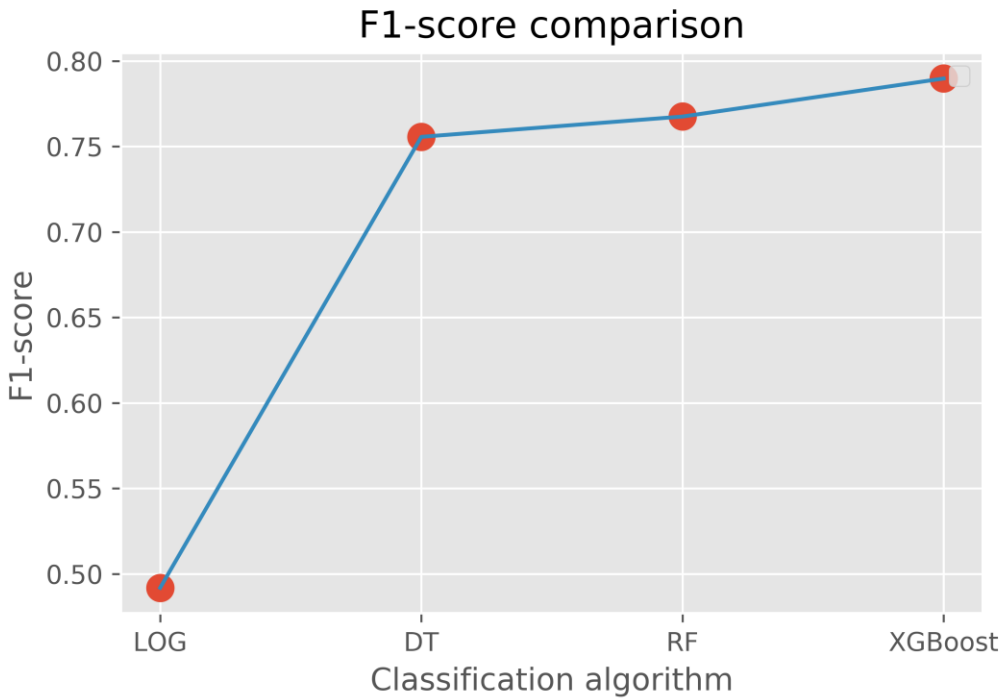

Random Forest
test-data
Predictions
Purchase
20000
15000
10000
5000
0
0
20
40
60
80
100

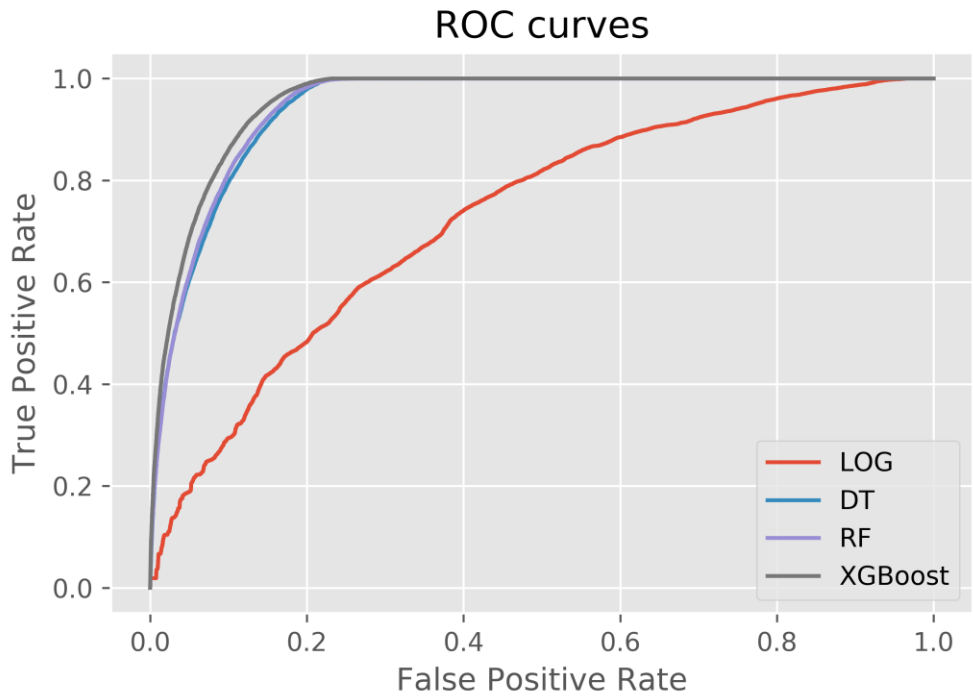

Multiple Linear Regression
test-data
Predictions
Purchase
20000
15000
10000
5000
0
0
20
40
60
80
100


Test Vs Predictions for various regression models
Purchase
20000
15000
10000
5000
0
0
20
40
60
80
100
test-data
MLR
DT
RF
XGBoost


R2-score comparison
R2-score
0.7
0.6
0.5
0.4
0.3
MLR
DT
RF
XGBoost
Regression algorithm


RMSE comparison
RMSE
4500
4000
3500
3000
2500
2000
train
test
MLR
DT
RF
XGBoost
Regression algorithm


F1-score comparison
0.80
0.75
0.70
0.65
0.60
0.55
0.50
F1-score
LOG
DT
RF
XGBoost
Classification algorithm


ROC curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
LOG
DT
RF
XGBoost