Digital Sales Prediction

Data Preprocessing & Feature Engineering

- Changing Down payment data from text to float
- Extracting only year from “Year of Birth” field and calculating approx. age
- Identifying gender from names (based on titles)
- Categorizing age based on generation – Millennials (22-37), Gen X(38-53), Baby boomers (53+)
- Categorizing CustomerType based on Previous Products – New (0), Engaged(1-3), Loyal(3+)

Model Selection (Supervised Learning)

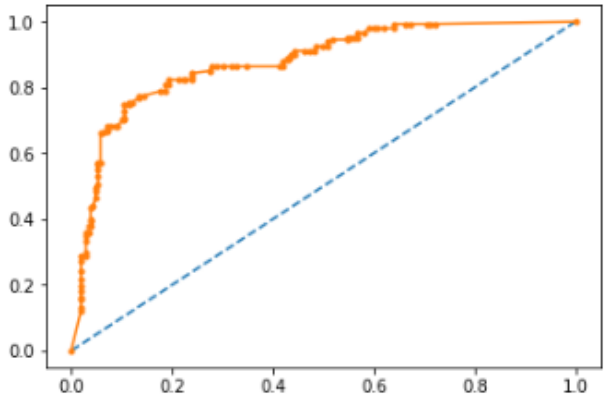
- Since it's a classification problem, I used Logistic Regression as a base model and compared its performance with RandomForest, DecisionTree, SVM, Adaboost and XGBoost. Following is the performance:

Model Name	Precision	Recall	F1 - Score	AUC
Logistic Regression	0.74	0.73	0.71	0.80
RandomForest	0.83	0.83	0.82	0.88
DecisionTree	0.82	0.82	0.82	0.88
SVM	0.75	0.75	0.74	0.80
XGBoost	0.76	0.76	0.75	0.83

Best Model Performance

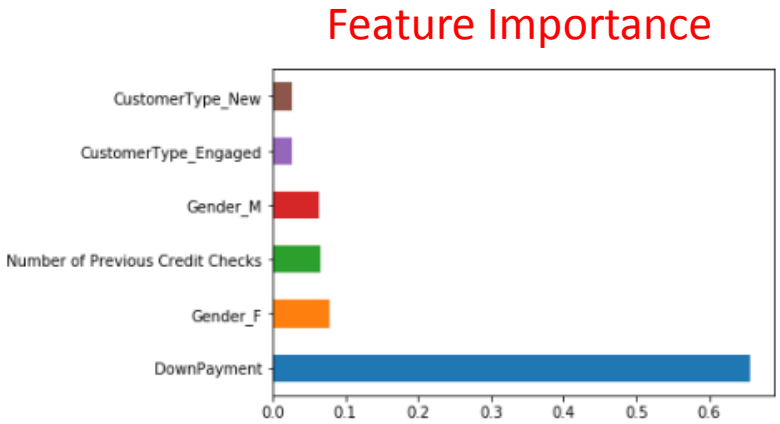
- RandomForest was the best performing model with the highest AUC of 0.88 and a precision and recall of 0.82. The AUC of 0.88 means that there is 88% chance of successful prediction of digital sales

AUC



Feature Importance

- Down payment is the biggest factor in determining the likelihood of digital sales
- Gender and previous credit checks are also importan



Unsupervised Learning (Kmeans Clustering)

- Performed Kmeans clustering to identify which clusters are more likely to make digital sales. Following is the summary:

Cluster	Sale	#Cstmrs	CC	DP	F	M	O	GenX	GenMil	GenBB	C_Eng	C_Loy	C_New
0	0.27	996	0.10	9.48	0.19	0.80	0.01	0.49	0.48	0.03	0.14	0.00	0.86
1	**0.65**	**34**	**1.18**	**238.19**	**0.47**	**0.53**	**0.00**	**0.53**	**0.47**	**0.00**	**0.41**	**0.00**	**0.59**
2	**0.65**	**130**	**0.51**	**78.45**	**0.28**	**0.65**	**0.08**	**0.60**	**0.35**	**0.05**	**0.52**	**0.11**	**0.37**
3	0.46	398	0.81	26.74	0.20	0.75	0.05	0.61	0.17	0.22	0.46	0.06	0.48
4	**1.00**	**6**	**0.33**	**512.33**	**0.33**	**0.67**	**0.00**	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**1.00**
5	**0.78**	**36**	**0.67**	**146.25**	**0.44**	**0.50**	**0.06**	**0.72**	**0.17**	**0.11**	**0.44**	**0.00**	**0.56**
6	0.51	144	0.67	50.03	0.13	0.85	0.03	0.64	0.25	0.11	0.44	0.14	0.42
7	**0.73**	**30**	**0.93**	**112.93**	**0.33**	**0.67**	**0.00**	**0.60**	**0.40**	**0.00**	**0.73**	**0.00**	**0.27**

Indicator	Mnemonics
Sale	Sale
CreditChecks	CC
DownPayment	DP
Gender_F	F
Gender_M	M
Gender_O	O
Generation_BabyBoomers	GenX
Generation_Gen	GenMil
Generation_Millenials	GenBB
CustomerType_Engaged	C_Eng
CustomerType_Loyal	C_Loy
CustomerType_New	C_New

- Digital Sales for Clusters 1,2,4,5 and 6 was higher. These clusters have people who paid more in Down payment, lot of customers belong to Gen X(38-54 yrs) and have not bought any products from RBC. This can be a good target market.