Project C :

Coronary Heart Disease Predictor

With Azure Deployment

Owner: Reema Lad

IMS ID No. : IMSPRO143500

Monday, May 19, 2021

Project Case

Problem Statement

A healthcare organization together with a couple of government hospitals in a city has collected information about the vitals that would reveal if the person might have a coronary heart disease in the next ten years or not.

Case Study Background

This study is useful in early identification of disease and have medical intervention if necessary. This would help not only in improving the health conditions but also the economy as it has been identified that health performance and economic performance are interlinked.

Understanding Data And EDA

Raw Data

Number of variables	25	Numeric	17
Number of observations	34281	Categorical	8
Missing cells	1743		
Missing cells (%)	0.2%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		

	ID	IV	A1	A2	A3	A4	A5	A6	A7	A8	...	A14	A15	A16	A17	A18	A19	A20	A21	A22	Target
0	1443894	2049	44	8.0	11	0	0	0	0	38	...	0	0.52	0.69	0	0	0	1	17.078971	0	0
1	1810849	48	0	8.0	0	0	0	0	0	1	...	0	0.59	0.78	1	0	0	1	17.022384	0	0
2	2264999	318	2	9.0	0	0	0	0	0	0	...	0	0.94	0.79	1	0	0	0	17.024773	0	0
3	1931676	62	4	2.0	0	0	0	15	30	7	...	0	0.51	0.47	0	0	0	1	17.074995	0	0
4	2070885	2	0	8.0	0	0	0	0	1	0	...	0	0.82	0.81	0	0	0	1	17.072697	0	0
5	1566137	6648	2327	6.0	1404	0	11604	23532	35880	10516	...	0	0.74	0.50	0	0	0	1	17.073619	0	0

Raw Data

Preprocessing

ID
Real number ($\mathbb{R}_{\geq 0}$)
UNIQUE

A2
Real number ($\mathbb{R}_{\geq 0}$)
MISSING
Missing 1743
Missing (%) 5.1%

A11
Categorical
CONSTANT

Target
Categorical
HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0

0	22988
1	11293

Preprocess

Missing Value Imputation

Missing Data may be represented by either NAs, Blanks or values such as -999/-99 etc.

A2 Variable has 1743 NAs

IV, A15 & A16 Variable's have 971, 2233, 2103 negative values, respectively.

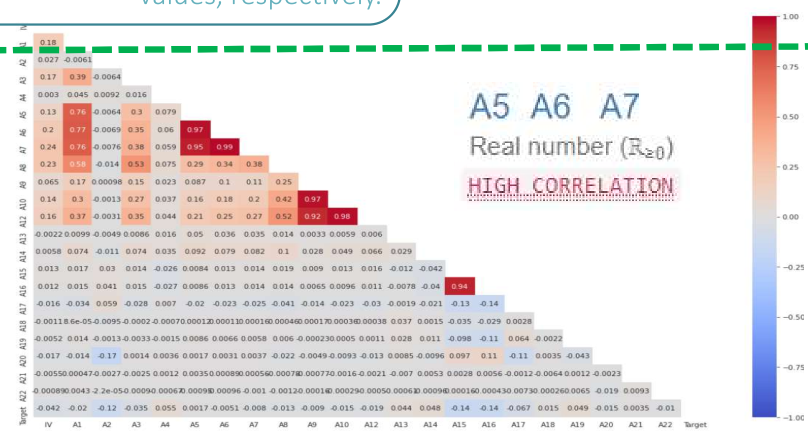
Replaced all negative values to "NAs", then imputed all "NAs" with "Mean"

Missing Value

Correlation Analysis

A15 A16
Real number ($\mathbb{R}$)
HIGH CORRELATION

A9 A10 A12
Real number ($\mathbb{R}_{\geq 0}$)
HIGH CORRELATION



A5 A6 A7
Real number ($\mathbb{R}_{\geq 0}$)
HIGH CORRELATION

Correlation

EDA And Data Preprocessing : Python

Original Data

ID	0
IV	0
A1	0
A2	1743
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10	0
A11	0
A12	0
A13	0
A14	0
A15	0
A16	0
A17	0
A18	0
A19	0
A20	0
A21	0
A22	0
Target	0

Data Post Imputing Negative Values

ID	0
IV	971
A1	0
A2	1743
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10	0
A11	0
A12	0
A13	0
A14	0
A15	2233
A16	2103
A17	0
A18	0
A19	0
A20	0
A21	0
A22	0
Target	0

Data Post Manipulation

ID	0
IV	0
A1	0
A2	0
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10	0
A11	0
A12	0
A13	0
A14	0
A15	0
A16	0
A17	0
A18	0
A19	0
A20	0
A21	0
A22	0
Target	0

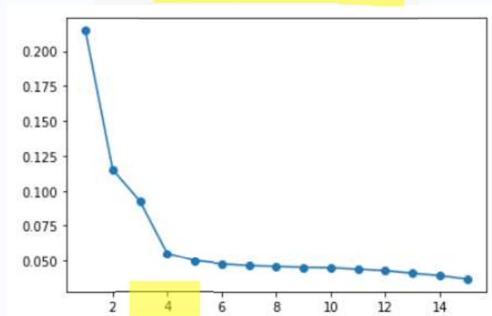
Data Post Cleaning

```
RangeIndex: 34281 entries, 0 to 34280
Data columns (total 23 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   IV      34281 non-null  float64
 1   A1      34281 non-null  float64
 2   A2      34281 non-null  float64
 3   A3      34281 non-null  float64
 4   A4      34281 non-null  float64
 5   A5      34281 non-null  float64
 6   A6      34281 non-null  float64
 7   A7      34281 non-null  float64
 8   A8      34281 non-null  float64
 9   A9      34281 non-null  float64
 10  A10     34281 non-null  float64
 11  A12     34281 non-null  float64
 12  A13     34281 non-null  float64
 13  A14     34281 non-null  float64
 14  A15     34281 non-null  float64
 15  A16     34281 non-null  float64
 16  A17     34281 non-null  float64
 17  A18     34281 non-null  float64
 18  A19     34281 non-null  float64
 19  A20     34281 non-null  float64
 20  A21     34281 non-null  float64
 21  A22     34281 non-null  float64
 22  Target  34281 non-null  float64
dtypes: float64(23)
```

Data Post PCA

```
RangeIndex: 34281 entries, 0 to 34280
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   PC_1    34281 non-null  float64
 1   PC_2    34281 non-null  float64
 2   PC_3    34281 non-null  float64
 3   PC_4    34281 non-null  float64
 4   Target  34281 non-null  float64
dtypes: float64(5)
```

```
PCA(n_components = 4)
```



Data Preprocessing Summary & Observations

Missing Value

Imputation of all “NAs”, <0 values, with Mean for : A2, IV, A15 & A16

Dropping Features

ID : as it’s a unique identity for each row AND A11 : as it has a constant value

High Corelation in Data

There is high corelation between : >>> A5, A6, A7 >>> A9, A10, A12 >>> A15, A16

Target Variable

Imbalanced Target Variable – “Target” >>> 0 : No : 22,988 >>> 1 : Yes : 11,293

Basis pre-processing, identified optimal **4 PC’s**, models will be deployed also using 4 PC’s along with all predictor variables

Predictor Variables

The data range of most features are too wide also there are outliers

Data Challenges & Resolution

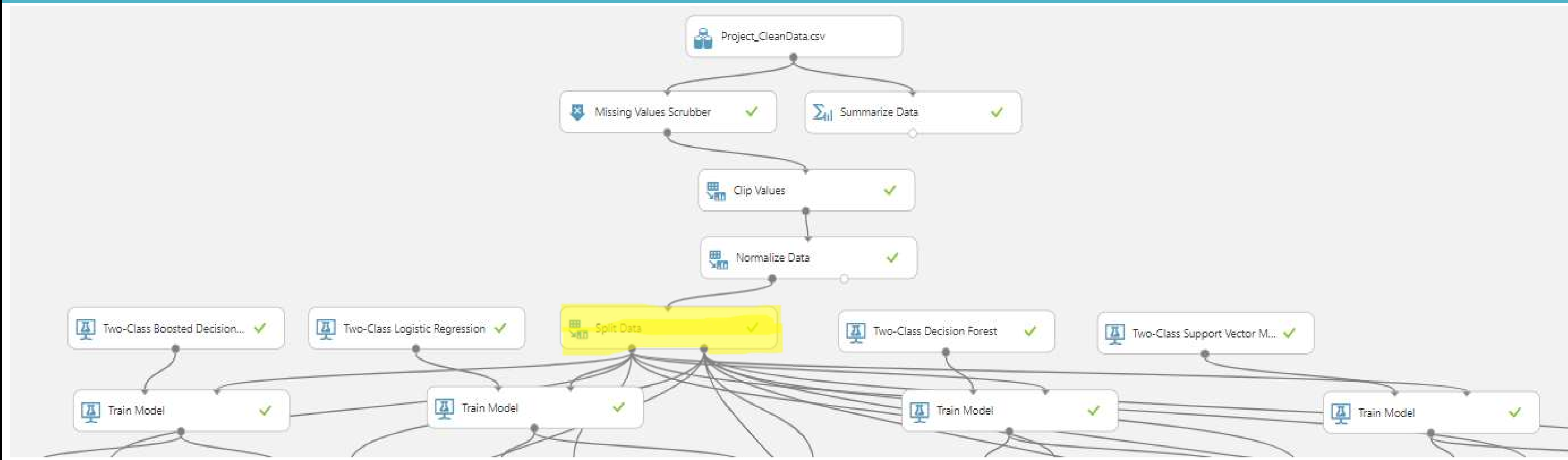
- As there is high corelation between certain predictor variables and the data is without clear feature definition / headers (explanation) hence would be assumed as unsupervised data, so attempted to create PC's and check the model performance using these components.

- Proposing to use Logistic, Random Forest (Bagging), GBM (Boosting). Will deploy using All Features as well as PC's, this might help in improving model performance.

- As Target variable is imbalanced will used stratified Split while deployment in Azure

- The predictor variables (most of them) have a very wide range as well as have outliers. For range variance in Azure pre-processed by Scaling and outliers treated with median value basis cut-off percentile

Azure Model : Train & Test



Target	Scored Labels	Scored Probabilities
0	0	0.245712
0	0	0.383242
0	0	0.171212
1	0	0.432151
0	0	0.489614
1	0	0.328744
0	0	0.411293
1	0	0.354784
1	0	0.465331
1	0	0.368748
0	0	0.324333
0	0	0.154041

Data Split : Cleaned Data 70-30%

Split Data › Results dataset1

rows	columns
23997	23

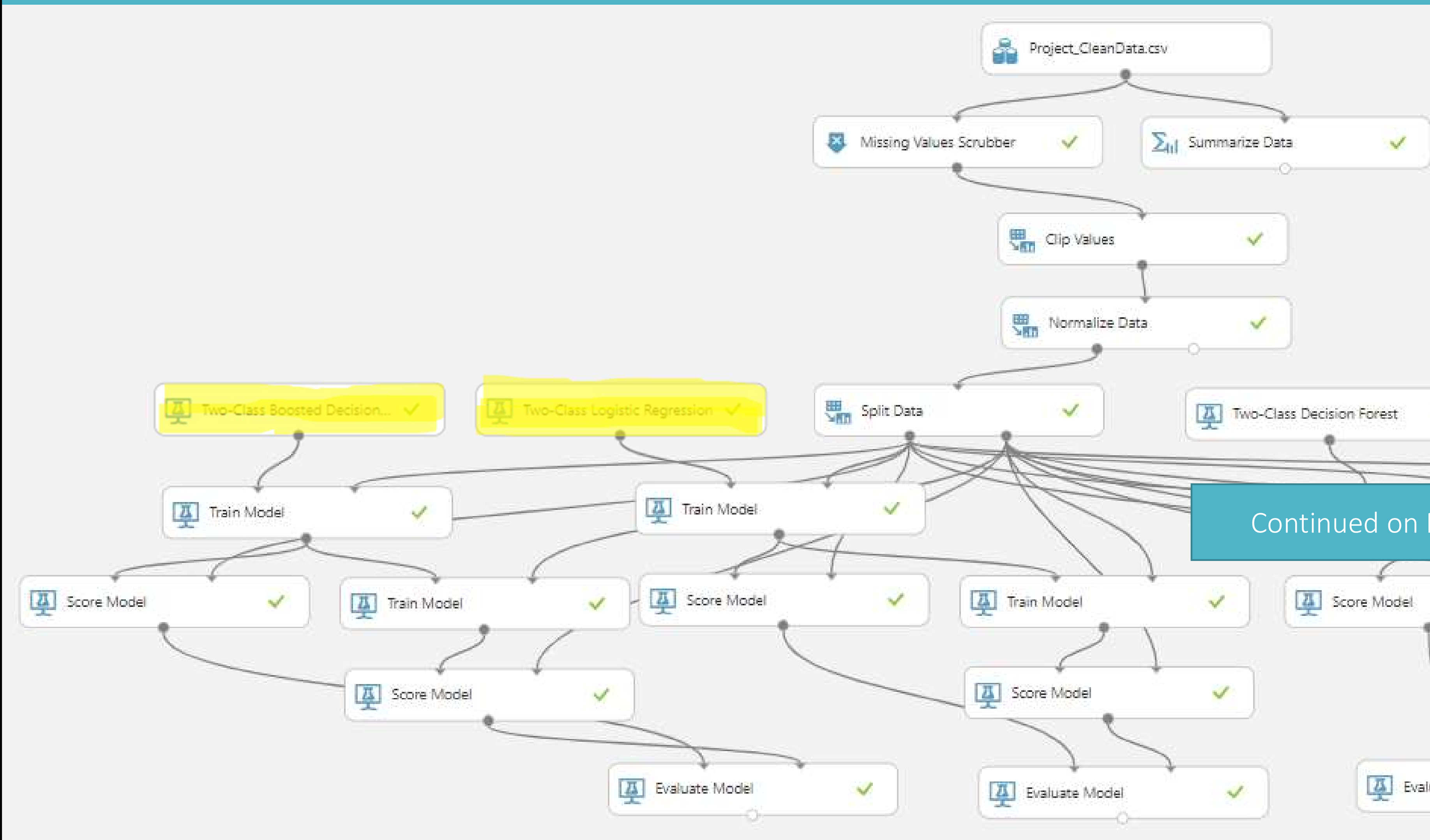
Split Data › Results dataset2

rows	columns
10284	23

Stratified split

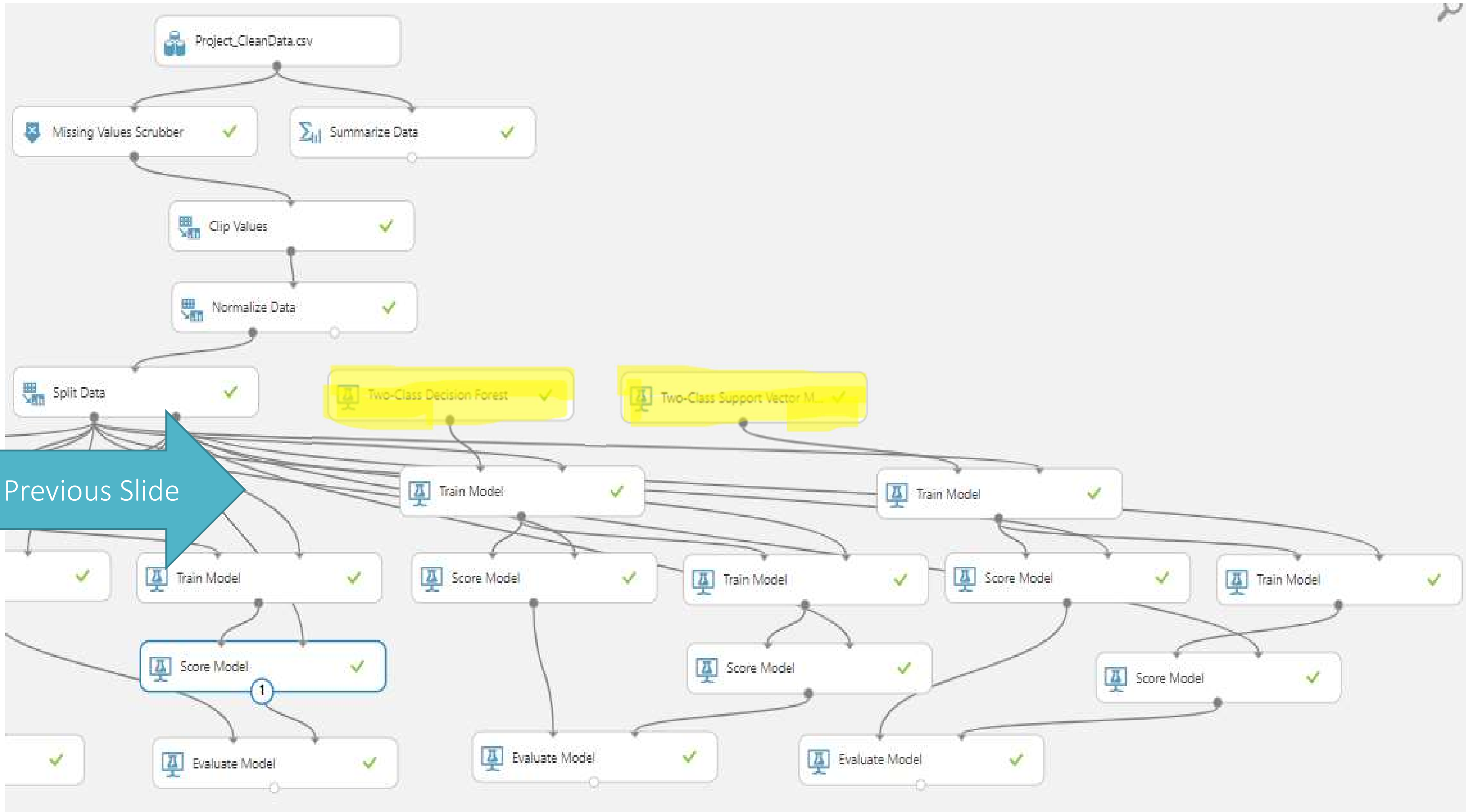
True

Azure Model Building : All Predictors



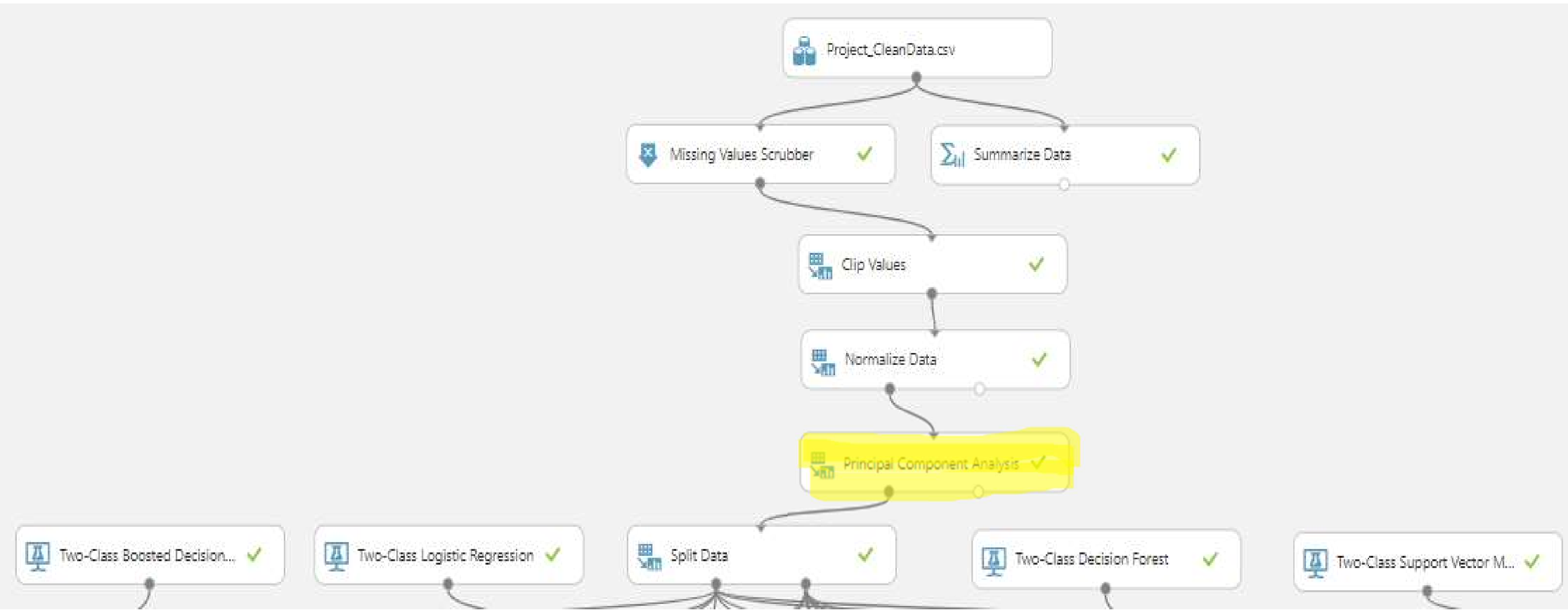
Continued on Next Slide

Azure Model Building : All Predictors



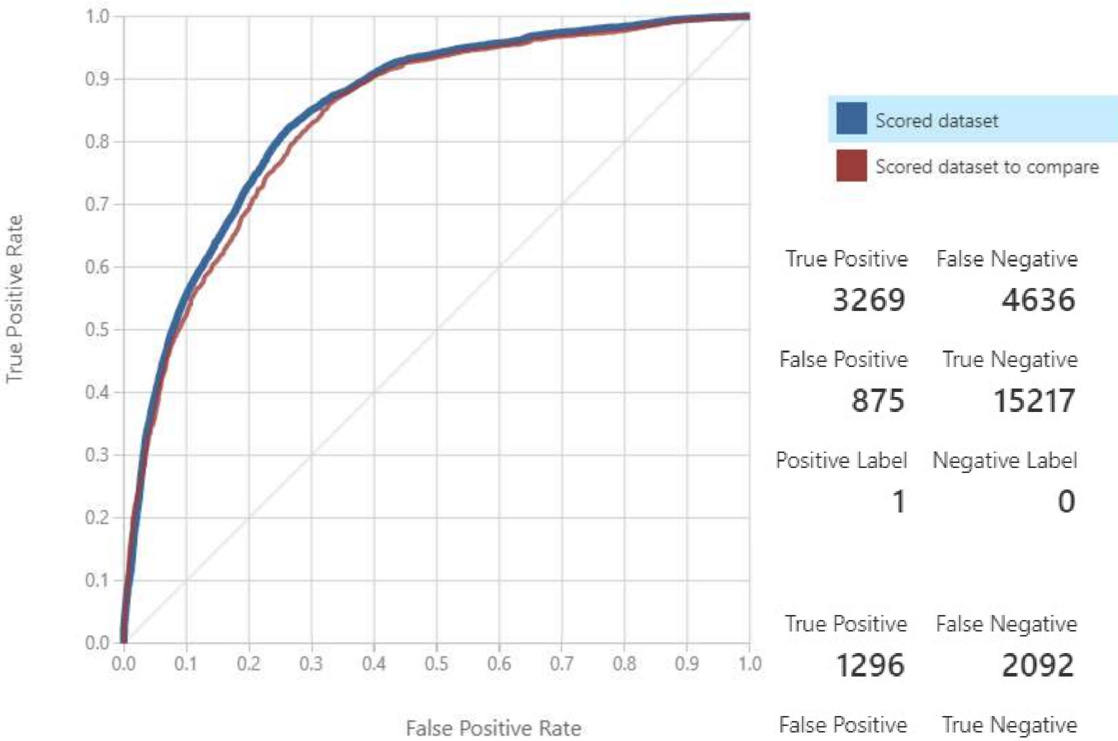
Continued from Previous Slide

Azure Model Building : Principal Components Only



Azure Model Building : Train & Test : Evaluation

Logistic Regression With All Features : Train – Test Evaluation



Train Evaluation

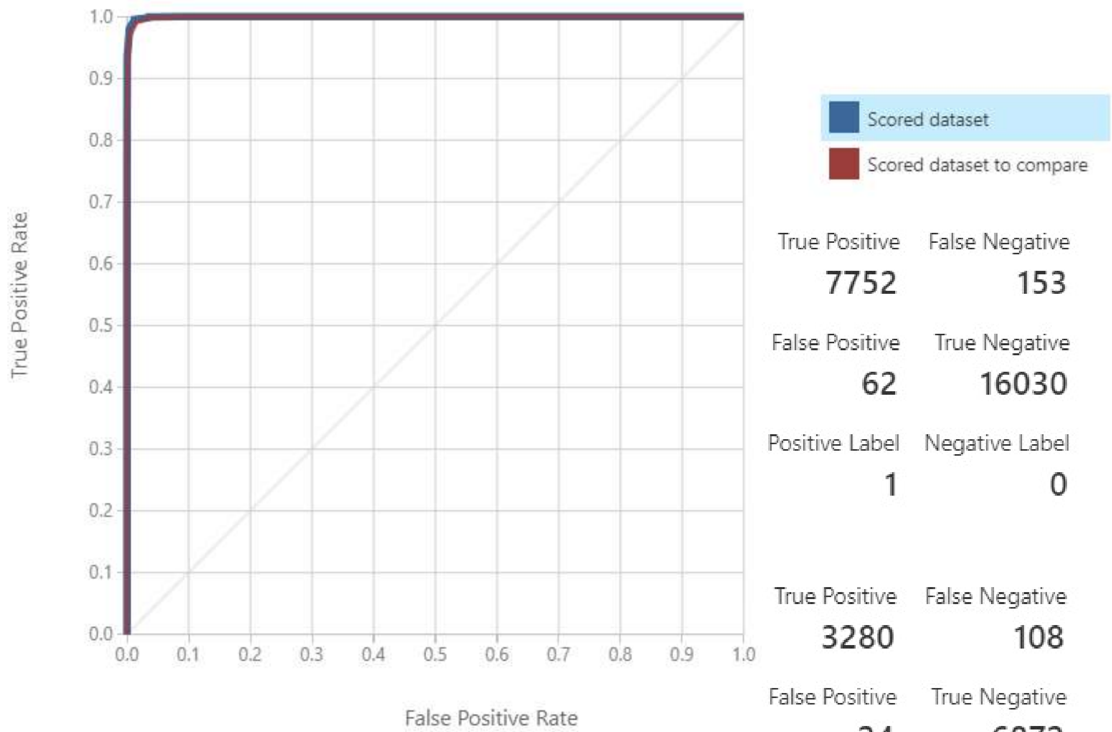
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3269	4636	0.770	0.789	0.5	0.849
False Positive	**True Negative**	**Recall**	**F1 Score**		
875	15217	0.414	0.543		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1296	2092	0.761	0.777	0.5	0.838
False Positive	**True Negative**	**Recall**	**F1 Score**		
371	6525	0.383	0.513		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

Random Forest With All Features : Train – Test Evaluation



Train Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
7752	153	0.991	0.992	0.5	1.000
False Positive	**True Negative**	**Recall**	**F1 Score**		
62	16030	0.981	0.986		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3280	108	0.987	0.993	0.5	0.999
False Positive	**True Negative**	**Recall**	**F1 Score**		
24	6872	0.968	0.980		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

Light GBM With All Features : Train – Test Evaluation

Train Evaluation

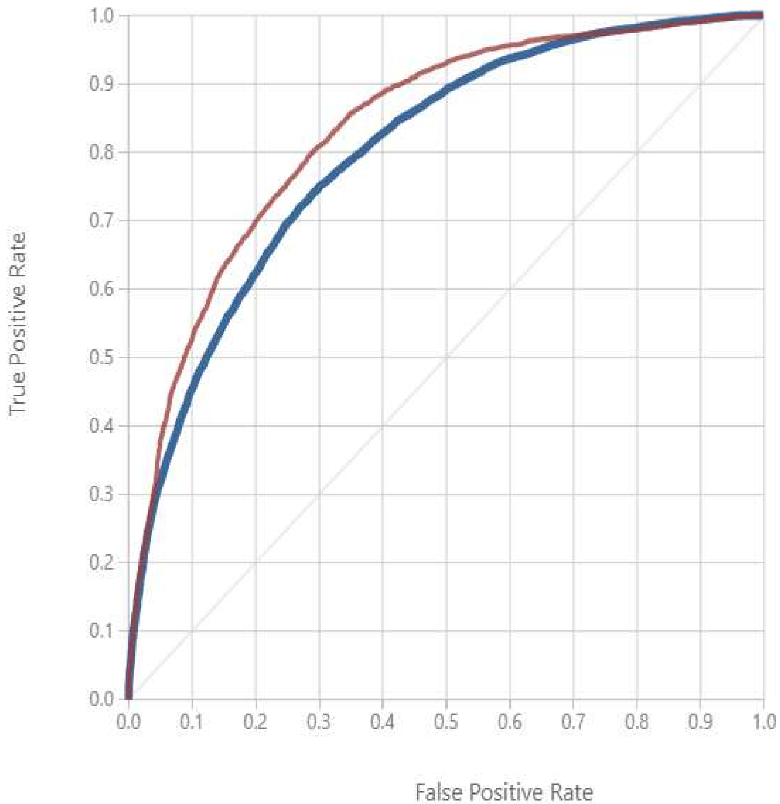
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
6771	1134	0.753	0.586	0.5	0.833
False Positive	**True Negative**	**Recall**	**F1 Score**		
4784	11308	0.857	0.696		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2867	521	0.745	0.577	0.5	0.831
False Positive	**True Negative**	**Recall**	**F1 Score**		
2101	4795	0.846	0.686		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

SVM With All Features : Train – Test Evaluation



Scored dataset

Scored dataset to compare

Train Evaluation

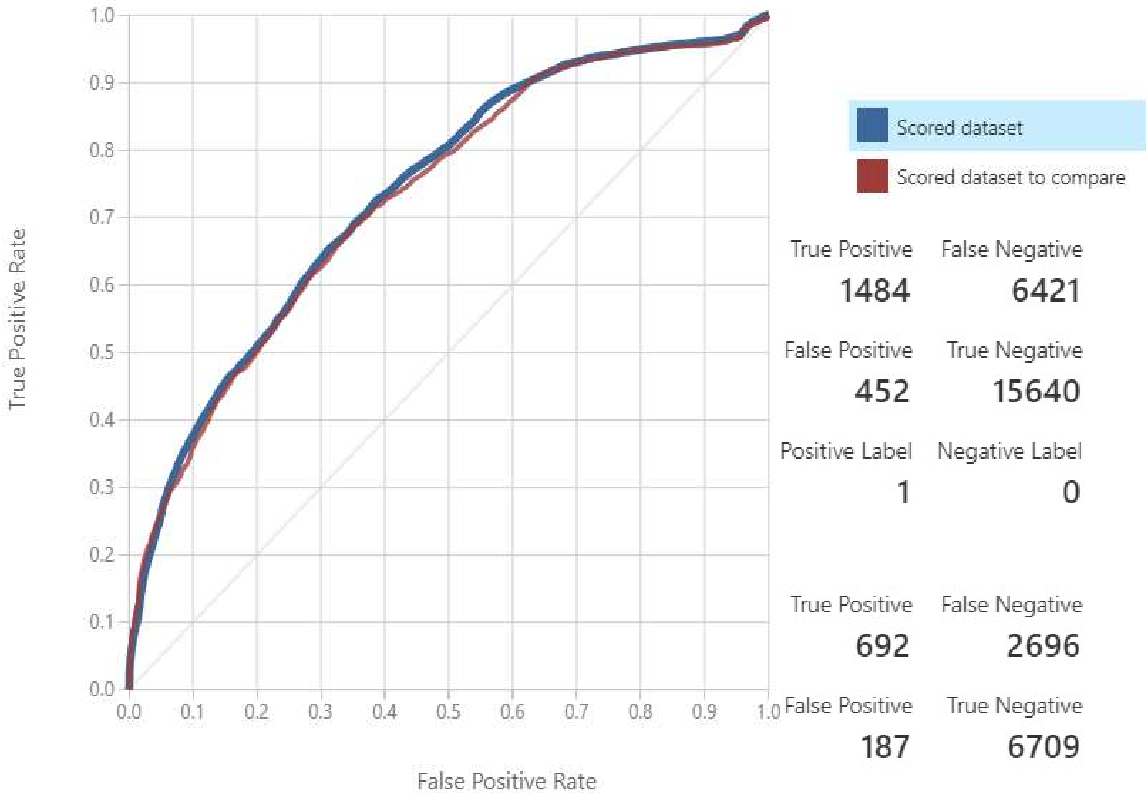
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3072	4833	0.748	0.718	0.5	0.801
False Positive	**True Negative**	**Recall**	**F1 Score**		
1204	14888	0.389	0.504		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1084	2304	0.747	0.784	0.5	0.834
False Positive	**True Negative**	**Recall**	**F1 Score**		
298	6598	0.320	0.455		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

Logistic Regression With PC's Only: Train – Test Evaluation



Train Evaluation

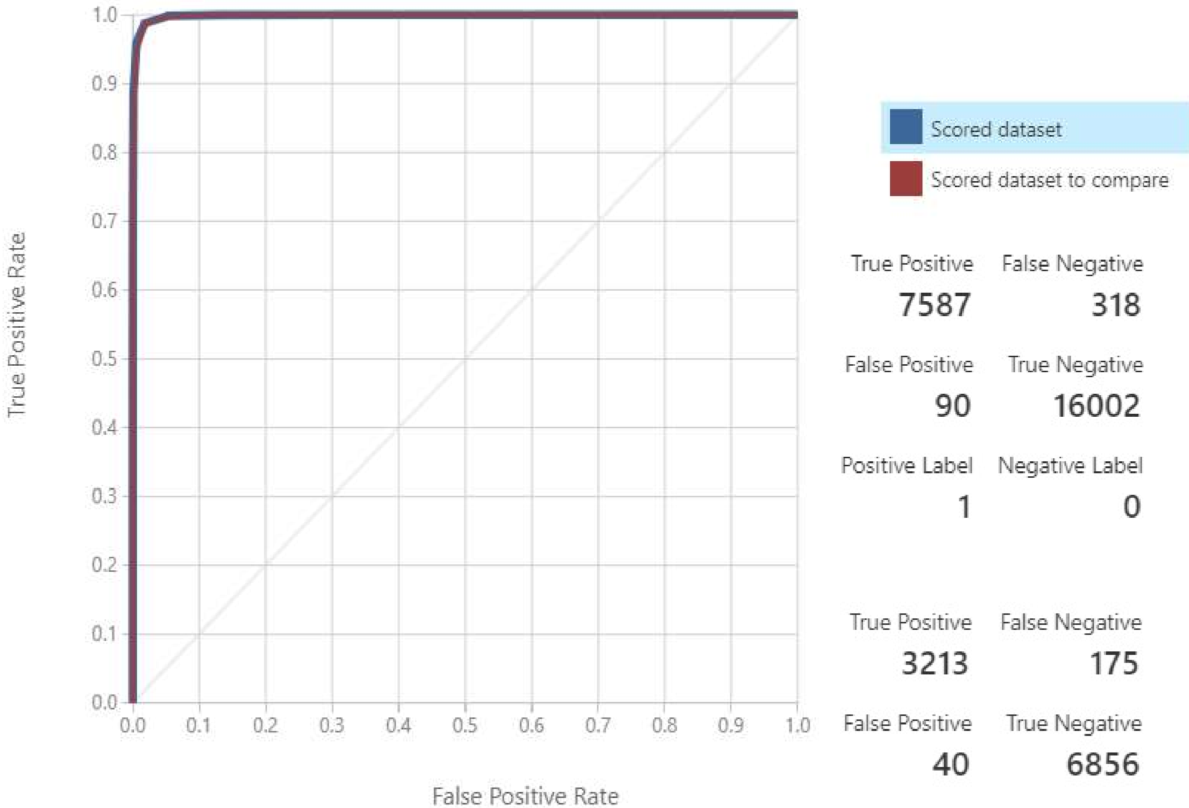
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1484	6421	0.714	0.767	0.5	0.736
False Positive	**True Negative**	**Recall**	**F1 Score**		
452	15640	0.188	0.302		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
692	2696	0.720	0.787	0.5	0.730
False Positive	**True Negative**	**Recall**	**F1 Score**		
187	6709	0.204	0.324		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

Random Forest With PC's Only: Train – Test Evaluation



Train Evaluation

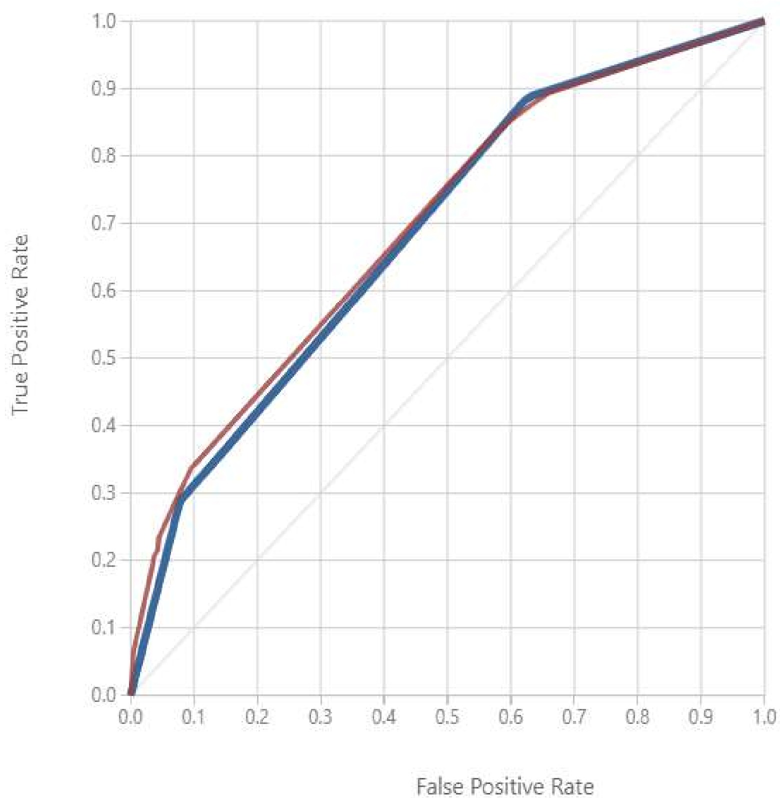
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
7587	318	0.983	0.988	0.5	0.999
False Positive	**True Negative**	**Recall**	**F1 Score**		
90	16002	0.960	0.974		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3213	175	0.979	0.988	0.5	0.999
False Positive	**True Negative**	**Recall**	**F1 Score**		
40	6856	0.948	0.968		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

Light GBM With PC's Only: Train – Test Evaluation



Scored dataset
Scored dataset to compare

Train Evaluation

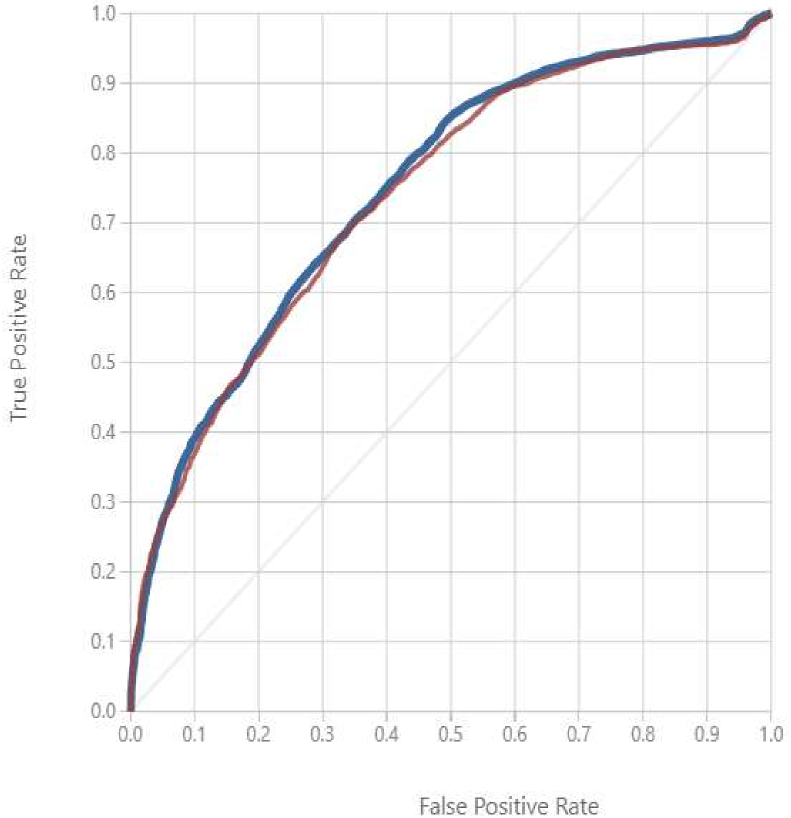
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2277	5628	0.713	0.644	0.5	0.686
False Positive	**True Negative**	**Recall**	**F1 Score**		
1257	14835	0.288	0.398		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1019	2369	0.718	0.658	0.5	0.697
False Positive	**True Negative**	**Recall**	**F1 Score**		
530	6366	0.301	0.413		
Positive Label	**Negative Label**				
1	0				

Azure Model Building : Train & Test : Evaluation

SVM With PC's Only: Train – Test Evaluation



Train Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1368	6537	0.711	0.776	0.5	0.747
False Positive	**True Negative**	**Recall**	**F1 Score**		
395	15697	0.173	0.283		
Positive Label	**Negative Label**				
1	0				

Test Evaluation

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
666	2722	0.719	0.800	0.5	0.739
False Positive	**True Negative**	**Recall**	**F1 Score**		
167	6729	0.197	0.316		
Positive Label	**Negative Label**				
1	0				

Azure Model : Evaluation

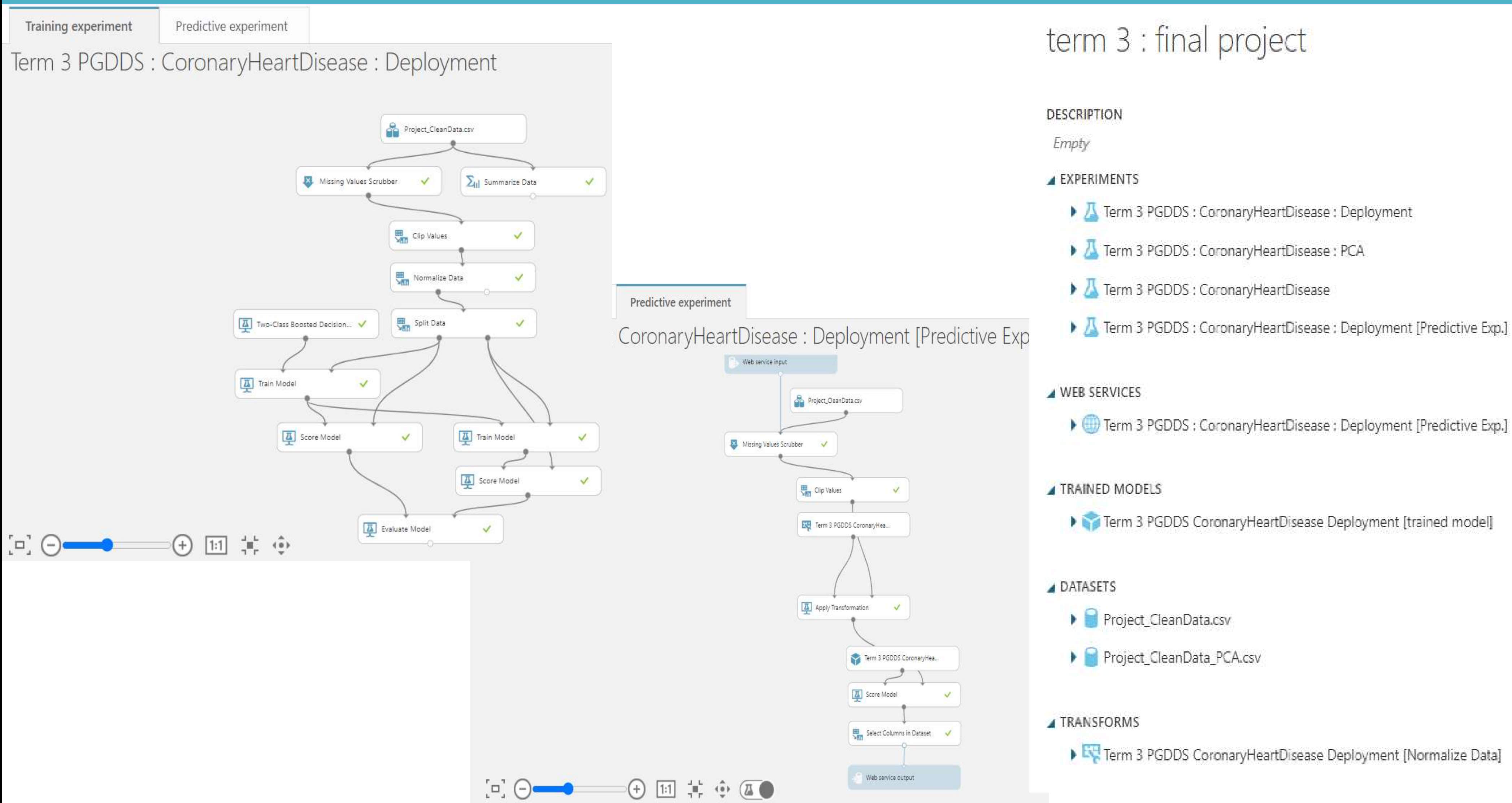
Model Name	True Positive	False Negative	False Positive	True Negative	Accuracy	Precision	Recall	F1 Score	Threshold	AUC	Positive Label	Negative Label
RF Test	3280	108	24	6872	0.987	0.993	0.968	0.98	0.5	0.999	1	0
RF Train	7752	153	62	16030	0.991	0.992	0.981	0.986	0.5	1	1	0
RF Test_PC	3213	175	40	6856	0.979	0.988	0.948	0.968	0.5	0.999	1	0
RF Train_PC	7587	318	90	16002	0.983	0.988	0.96	0.974	0.5	0.999	1	0
LGBM Test	2867	521	2101	4795	0.745	0.577	0.846	0.686	0.5	0.831	1	0
LGBM Train	6771	1134	4784	11308	0.753	0.586	0.857	0.696	0.5	0.833	1	0
LR Test	1296	2092	371	6525	0.761	0.777	0.383	0.513	0.5	0.838	1	0
SVM Test	1084	2304	298	6598	0.747	0.784	0.32	0.455	0.5	0.834	1	0
LGBM Test_PC	1019	2369	530	6366	0.718	0.658	0.301	0.413	0.5	0.697	1	0
LR Test_PC	692	2696	187	6709	0.72	0.787	0.204	0.324	0.5	0.73	1	0
SVM Test_PC	666	2722	167	6729	0.719	0.8	0.197	0.316	0.5	0.739	1	0
LR Train	3269	4636	875	15217	0.77	0.789	0.414	0.543	0.5	0.849	1	0
SVM Train	3072	4833	1204	14888	0.748	0.718	0.389	0.504	0.5	0.801	1	0
LGBM Train_PC	2277	5628	1257	14835	0.713	0.644	0.288	0.398	0.5	0.686	1	0
LR Train_PC	1484	6421	452	15640	0.714	0.767	0.188	0.302	0.5	0.736	1	0
SVM Train_PC	1368	6537	395	15697	0.711	0.776	0.173	0.283	0.5	0.747	1	0

Model Selection

Concluded to select Boosting Model, Light GBM, with all Features; siting the following reason :

- Highest accuracy is in Random Forest, using all Features as well as principal components, however considering the possibility of overfitting, considering the next best model.
- ‘False Negative’ score of each model is given the highest priority while selecting the model, as given the problem statement, where we want to identify the possibility of ‘Coronary Heart Disease’, Type 2 error needs to be avoided, hence minimal ‘False Negative’ predictor is considered, followed by model accuracy, consistency across Train & Test models and lastly AUC.

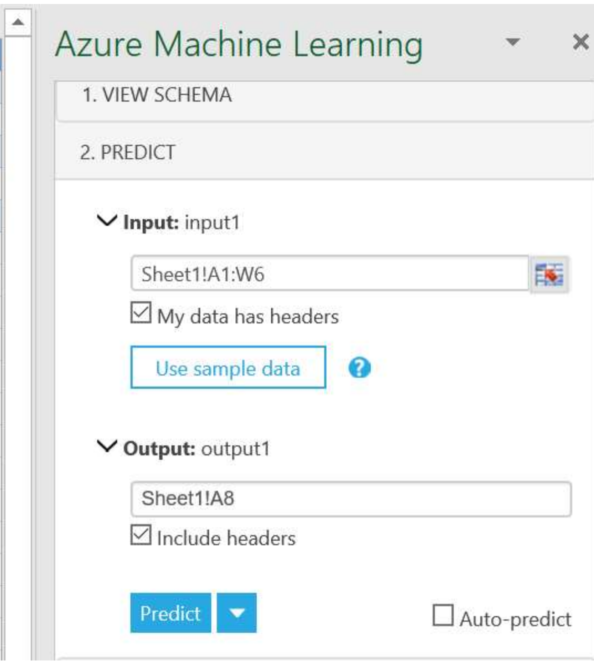
Azure Model Deployment : Predictive Experiment Deployment & Project Building in Azure



Azure Model Deployment : Predictive Experiment : Excel Prediction

IV	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
2049	44	8	11	0	0	0	0	38	123	
48	0	8	0	0	0	0	0	1	1	
318	2	9	0	0	0	0	0	0	1	
62	4	2	0	0	0	15	30	7	24	
2	0	8	0	0	0	0	1	0	0	

Target	Scored Labels	Scored Probabilities
0	0	0.078072175
0	0	0.078072175
0	0	0.078072175
0	0	0.113560125
0	0	0.078072175



Thank You