US Income Classification

Data set

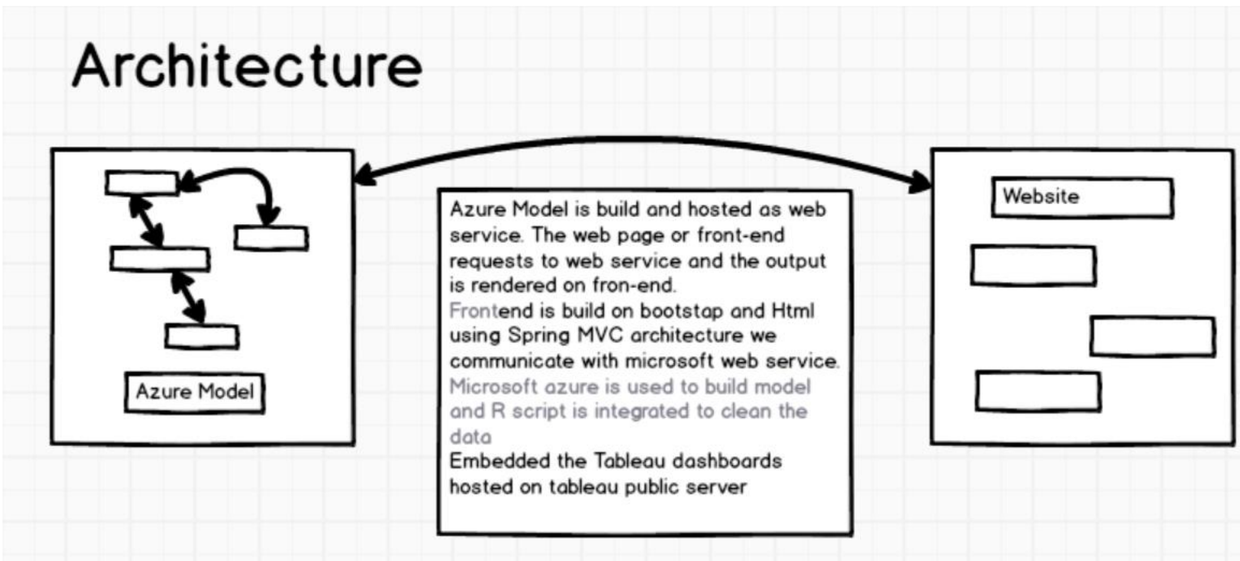
- In the census data every record represents a person with 14 attributes, the last element of a record is one of the labels {>=50K,<50K}.
- Age: continuous
- Workclass: 8 values
- Education: 16 values
- Education-num: continuous.
- Marital-status: 7 values
- Occupation: 14 values
- Relationship: 6 values
- Race: 5 values
- Sex: Male, Female
- Capital-gain: continuous.
- Capital-loss: continuous.
- Hours-per-week: continuous.
- Native-country: 41 values
- >50K Income: Yes, No

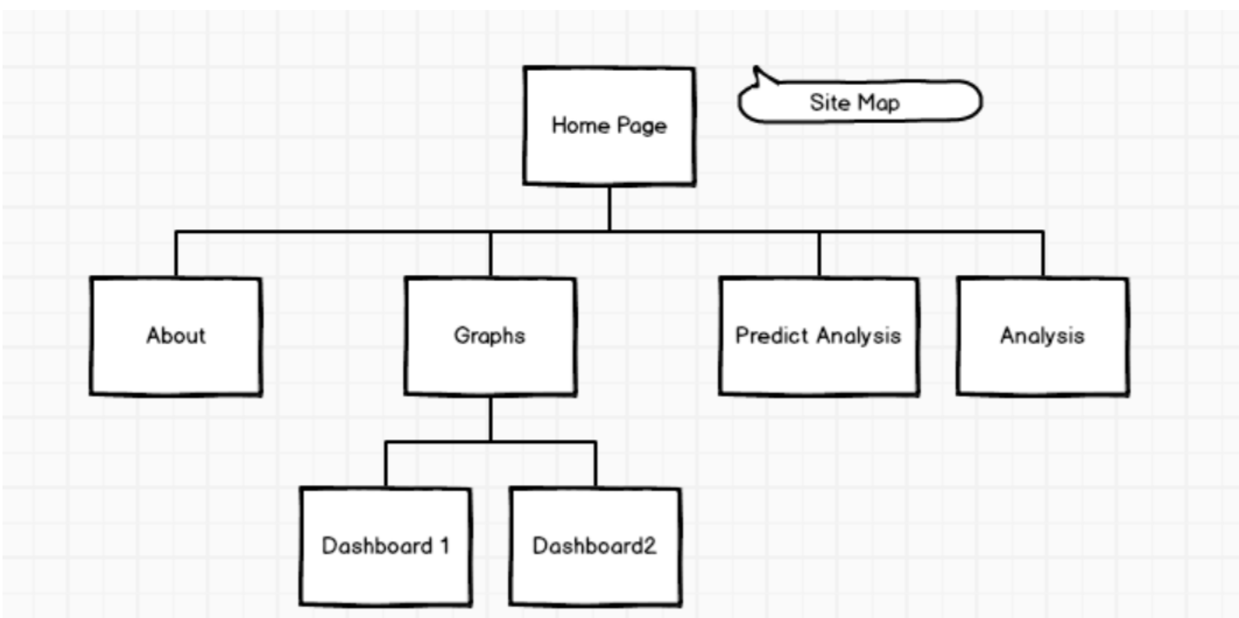
Business Problem

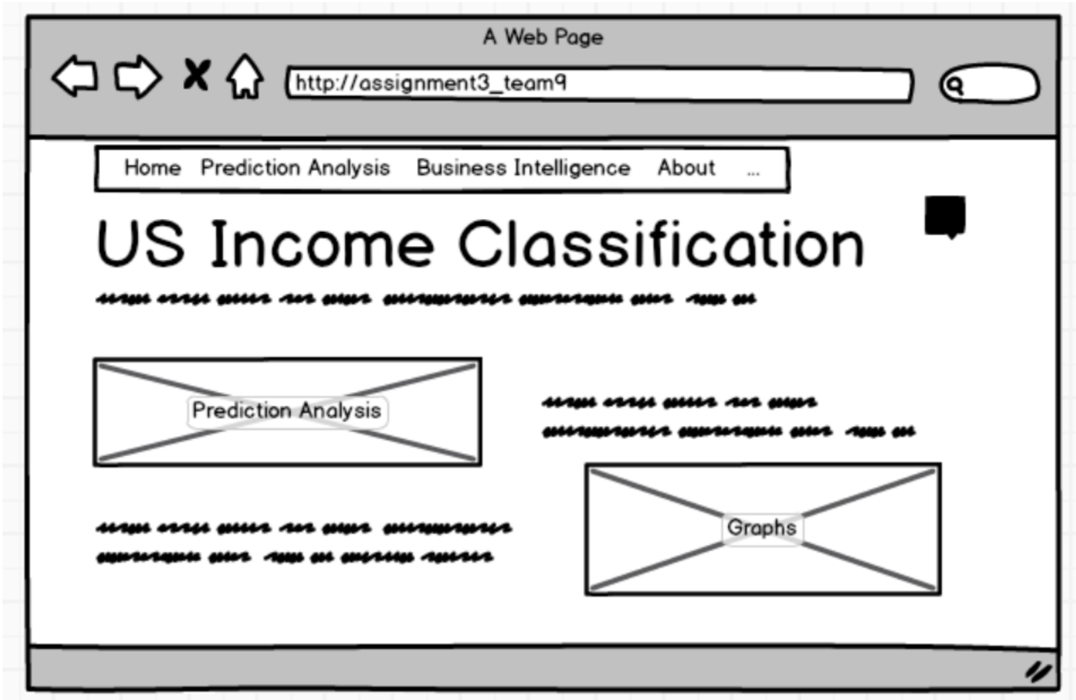
- Q1. Which of the variables (age, occupation, sex, etc.) are most decisive for determining the income of a person?
- Q2. Which values for which variables form conditions that would imply high income or low income?
- Q3. What percentage of people of which race, Occupation, Education, Age, status can afford what kind of Standard of living

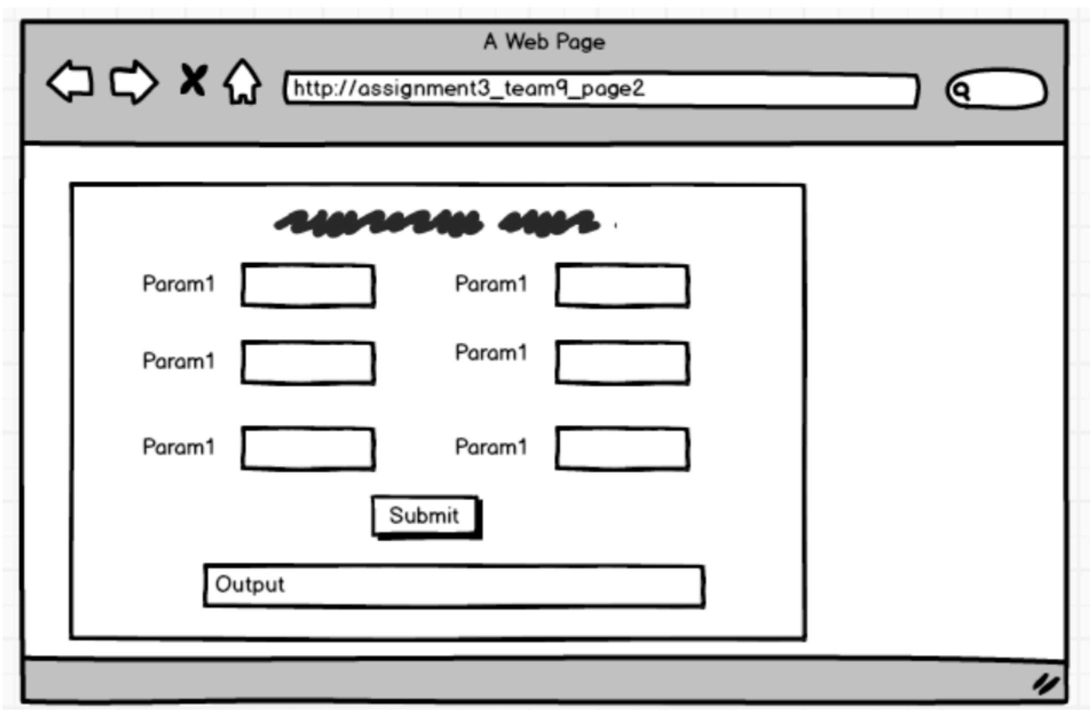
Patterns in Data Set

- **Age vs Job-** younger people tend to work in the private sector while older people work for the local government or are self employed
- **Age vs Number of Years In School**- older a person is, the more likely he/she is to have a greater number of years of education
- **50K vs School Level-** people who finish college are significantly more likely to earn over 50K






A Web Page
http://assignment3_team9
Home
Prediction Analysis
Business Intelligence
About
...
US Income Classification
Prediction Analysis
Graphs


A Web Page
http://assignment3_team9_page2
Param1
Param1
Param1
Param1
Param1
Param1
Submit
Output

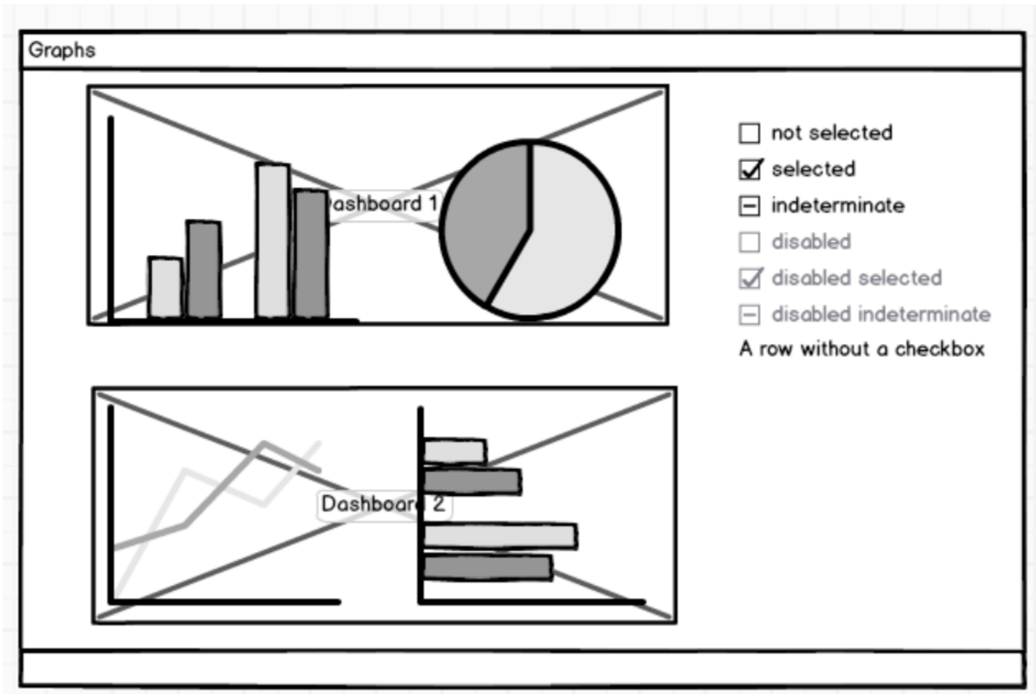
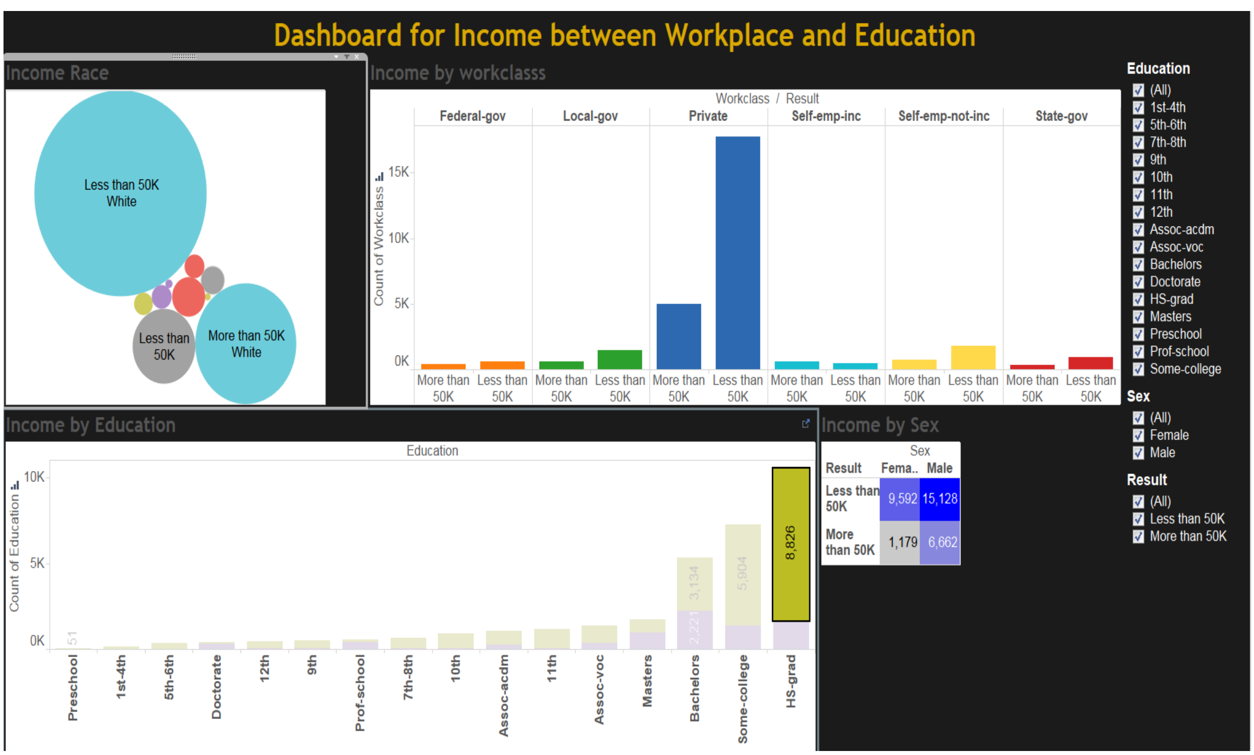
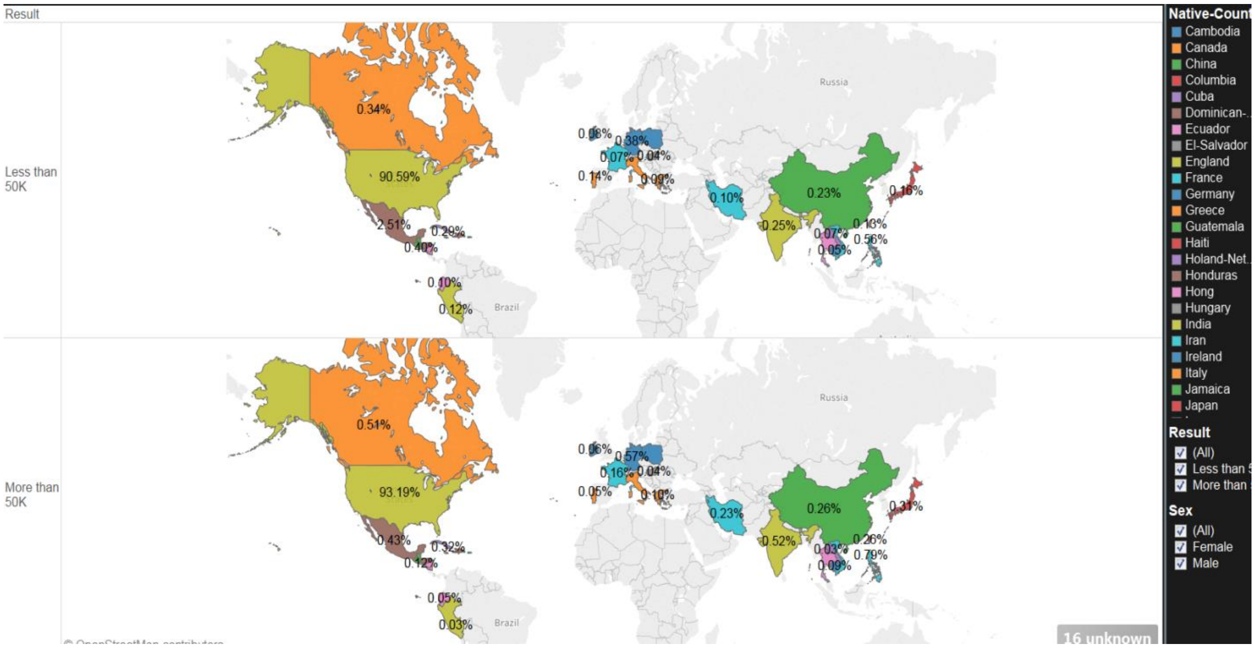


Tableau Visualization





Algorithms

We build following algorithms to solve above problem

1. Two - class boosted Decision Tree
2. Two - class Decision Jungle
3. Two - class Logistic Regression
4. Two – class Neural Network
5. Two – class Decision Forest

Designing of Algorithms in Azure and R

1. **Import the training dataset and include in Azure data set**
2. **Added R block to transform the data**
    - **Dataset contained invalid data in some columns with '?'**
    - **Replaced the invalid characters with NA's and converted them to factors**
3. **Selected columns from datasets which are functions required to calculate the income**

4. **Cleaned the missing data using Multivariate imputation using chained equations(MICE). Each variable with missing data is modeled conditionally using the other variable in the data before filling the missing values.**
5. **Added 3 metadata modules**
    I. **Selected income #result column as label to predict**
    II. **Selected String variables and made them categorial**
    III. **Selected other Numeric fields and setted its field as Integer**
6. **Filter based feature selection was used to select the important features which helps predicting the binary income classifier. Perarson's coorelation was as a base to score the important features**
7. **Used 5 different two class algorithm of which Two-class boosted decision algorithm resulted in highest accuracy of all.**
8. **Trained the model for the training dataset**
9. **Imported test dataset and process with steps 2 to 5**
10. **Scored the trained model for above test dataset**
11. **Evaluate the model for accuracy and use it as a web service.**
12. **The service was implemented in a web application to predict the income level when provided with the required input parameters**
13. **The web application was hosted in AWS at below url.**
    http://census-env-new.us-west-2.elasticbeanstalk.com/

ROC Curve

- All the algorithms were evaluated and Two-Class Boosted Tree was found to be best with 87% accuracy

- Two-class Boosted Algorithm was used to deploy the web service and classification analysis uses this model in the web application

Integration

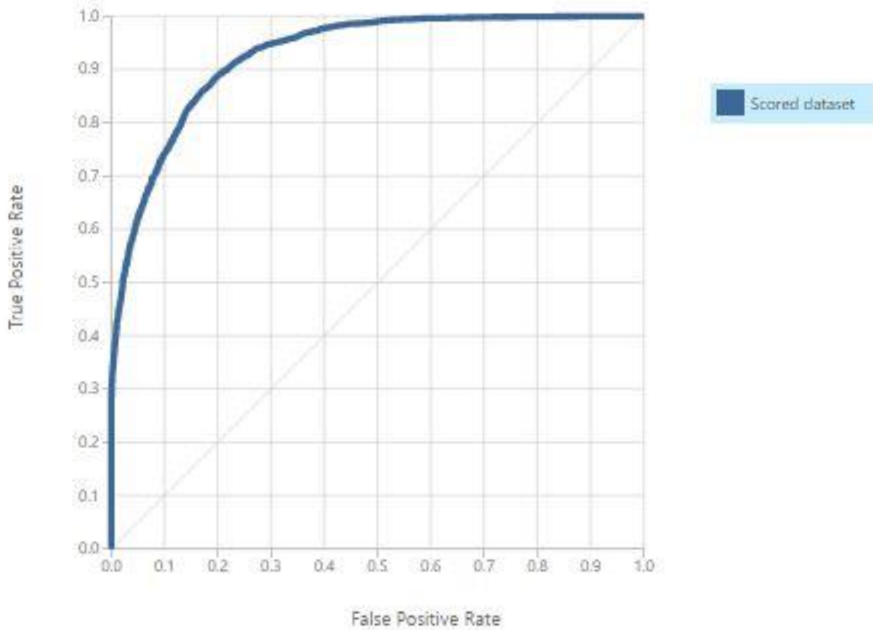
Azure

https://ussouthcentral.services.azureml.net/workspaces/855d6eb6f36e47eeaebfc6a8241e7cf2/services/a2159cb8885a4d94887ee0b04094baa7/execute?api-version=2.0&details=true
AWS domain:
http://census-env-new.us-west-2.elasticbeanstalk.com
Tableau Dashboard:

https://public.tableau.com/views/Assignment3_128/IncomebyNativeCountry?:embed=y&:display_count=yes
https://public.tableau.com/views/Assignment3_128/DashboardfprIncomebetweenWorkplaceandEducation?:embed=y&:display_count=yes

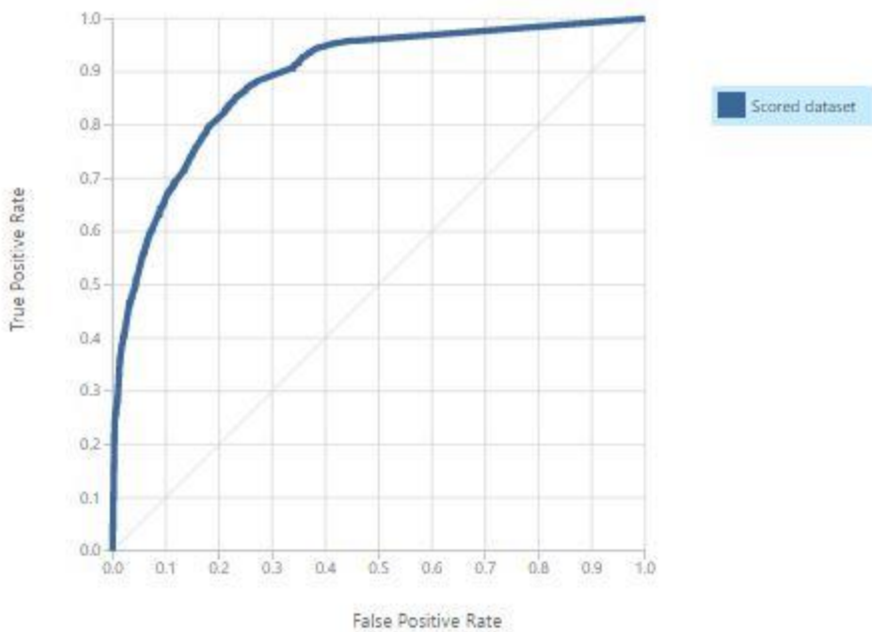
****

ROC PRECISION/RECALL LIFT



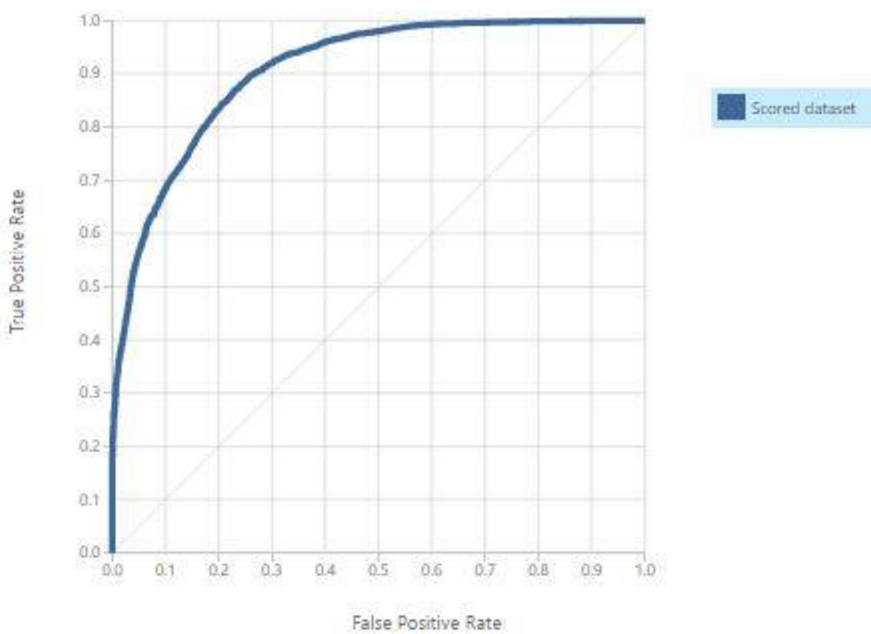
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2581	1265	0.870	0.751	0.5	0.927
False Positive	**True Negative**	**Recall**	**F1 Score**		
857	11577	0.671	0.709		
Positive Label	**Negative Label**				
>50K.	<=50K.				

ROC PRECISION/RECALL LIFT



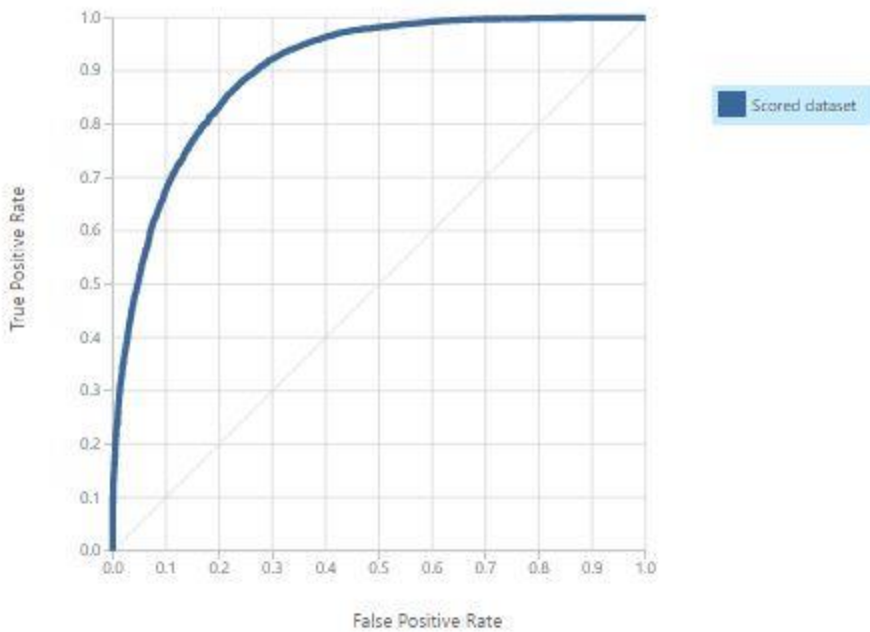
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2426	1420	0.847	0.695	0.5	0.888
False Positive	True Negative	Recall	F1 Score		
1066	11368	0.631	0.661		
Positive Label	Negative Label				
>50K.	<=50K.				

ROC PRECISION/RECALL LIFT



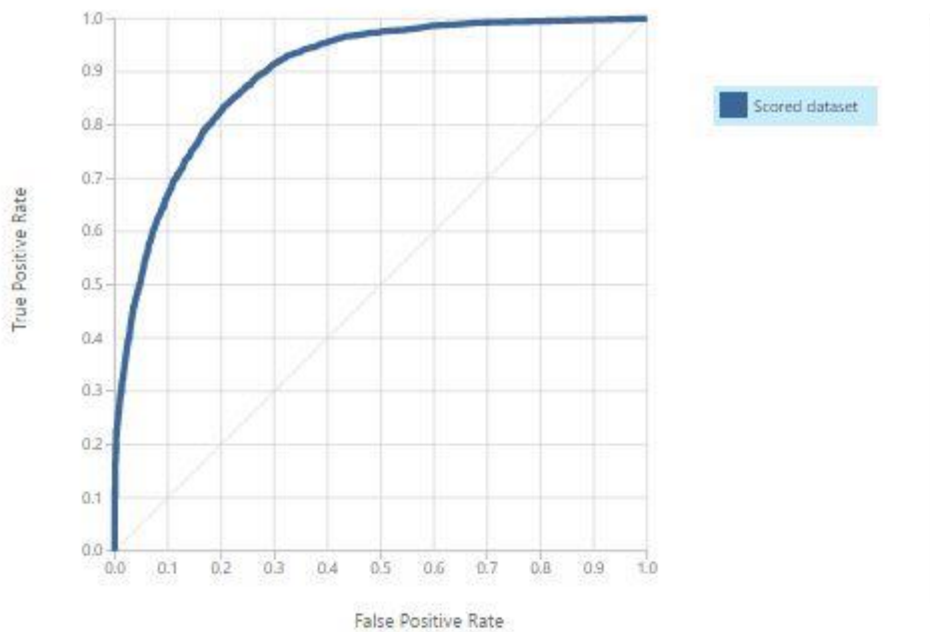
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2311	1535	0.858	0.749	0.5	0.906
False Positive	**True Negative**	**Recall**	**F1 Score**		
776	11658	0.601	0.667		
Positive Label	**Negative Label**				
>50K.	<=50K.				

ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2222	1624	0.849	0.728	0.5	0.903
False Positive	**True Negative**	**Recall**	**F1 Score**		
830	11604	0.578	0.644		
Positive Label	**Negative Label**				
>50K.	<=50K.				

ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2277	1569	0.850	0.722	0.5	0.898
False Positive	**True Negative**	**Recall**	**F1 Score**		
875	11559	0.592	0.651		
Positive Label	**Negative Label**				
>50K.	<=50K.				