Predicting if income exceeds $50,000 per year based on 1994 US Census Data

CAPSTONE PROJECT 2

PRINCE SEKYI

Problem Statement and Goals

Problem Statement

- Income inequality is the extent to which income is distributed. The distribution of income is very important for the economic growth of any country. Over the years this topic has been the concern of the government and other international organizations. In the United States, income inequality has been of so much concern and a lot of factors have been attributed to the uneven distribution of income.

Goals

This project will investigate

1. factors that affect the income level of an individual

2. Use machine learning to predict the income level of an individual

Data Cleaning and Data Wrangling

Below is the original data

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Below is the clean data

Weight	Education	Education_Num	Marital_Status	Occupation	Relationship	Race	Sex	Capital_gain	Capital_loss	Hours_per_week	Native_country	Income
77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

- **Missing values were removed**
- **Variable names were cleaned and written appropriately**

Exploratory Data Analysis

Dataset

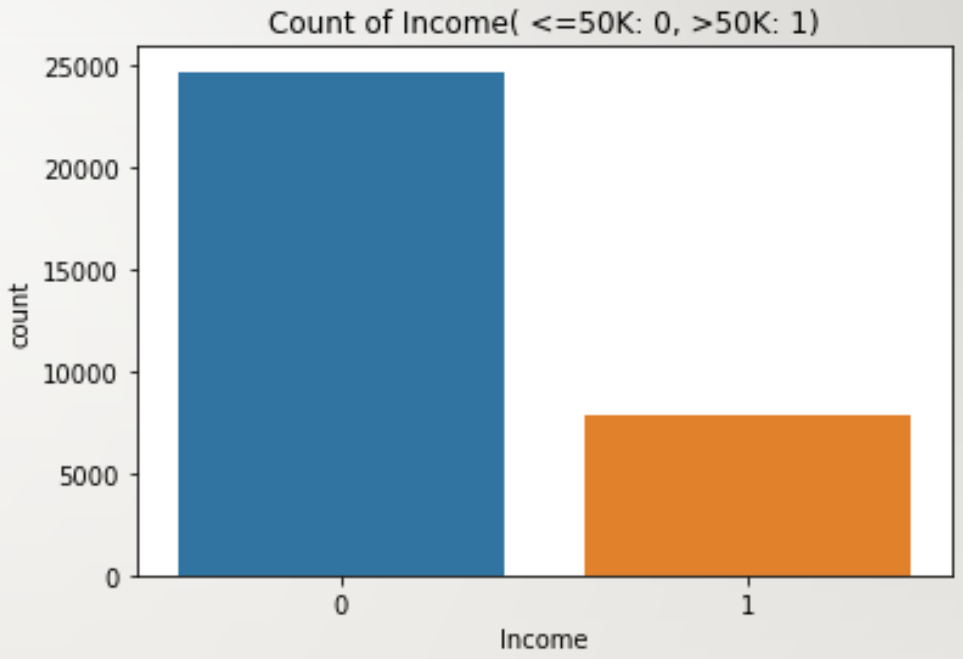
- **Age**: the age of an individual
- **Workclass**: a general term to represent the employment status of an individual
- **Fnlwgt**: final weight. In other words, this is the number of people the census believes the entry represents.
- **Education**: the highest level of education achieved by an individual.
- **Education_Num**: the highest level of education achieved in numerical form.
- **Marital_Status**: marital status of an individual.

Dataset

- **Occupation**: the general type of occupation of an individual
- **Relationship**: represents what this individual is relative to others.
- **Race**: Descriptions of an individual's race
- **Sex**: the biological sex of the individual
- **Capital_Gain**: capital gains for an individual
- **Capital_Loss**: capital loss for an individual
- **Hours_Per_Week**: the hours an individual has reported to work per week
- **nativecountry**: country of origin for an individual
- **Income**: whether or not an individual makes more than $50,000 annually

EDA contiunes

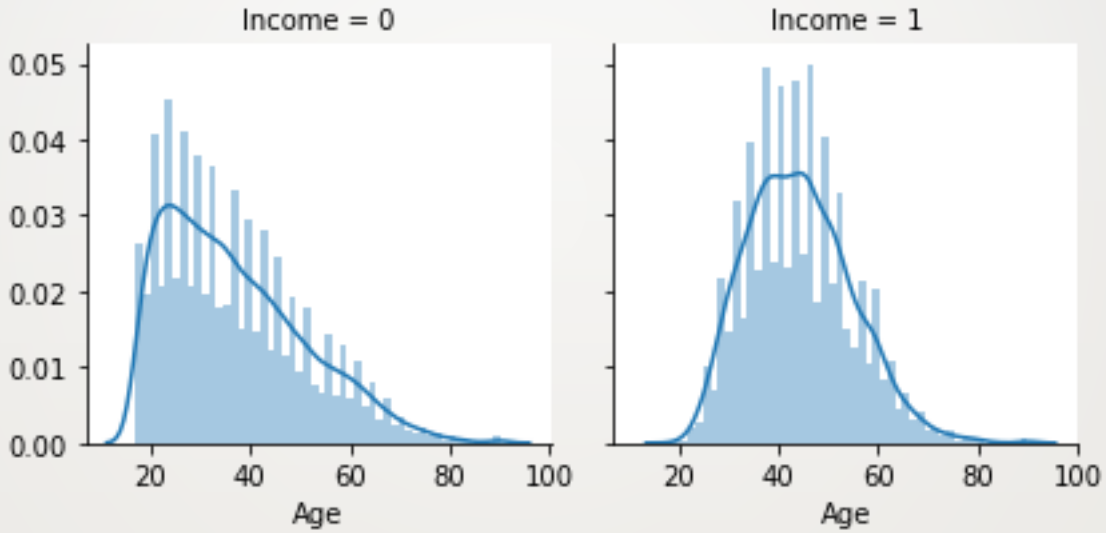
Label	Count
<=50K (0)	7841
>50K(1)	24720



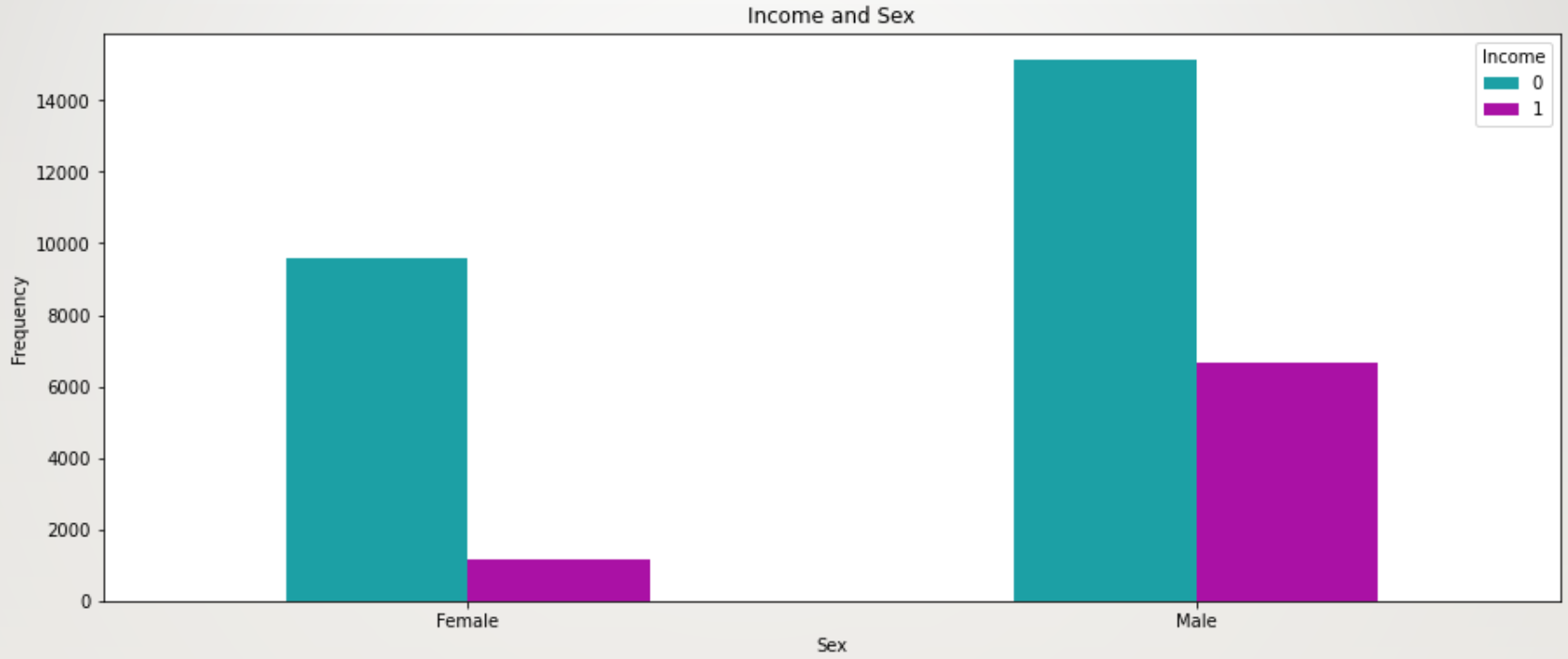
Heat Map


Age
Final_Weight
Education_Num
Capital_gain
Capital_loss
Hours_per_week
Income
1 -0.077 0.037 0.078 0.058 0.069 0.23
-0.077 1 -0.043 0.00043 -0.01 -0.019 -0.0095
0.037 -0.043 1 0.12 0.08 0.15 0.34
0.078 0.00043 0.12 1 -0.032 0.078 0.22
0.058 -0.01 0.08 -0.032 1 0.054 0.15
0.069 -0.019 0.15 0.078 0.054 1 0.23
0.23 -0.0095 0.34 0.22 0.15 0.23 1
Age
Final_Weight
Education_Num
Capital_gain
Capital_loss
Hours_per_week
Income
1.0
0.8
0.6
0.4
0.2
0.0

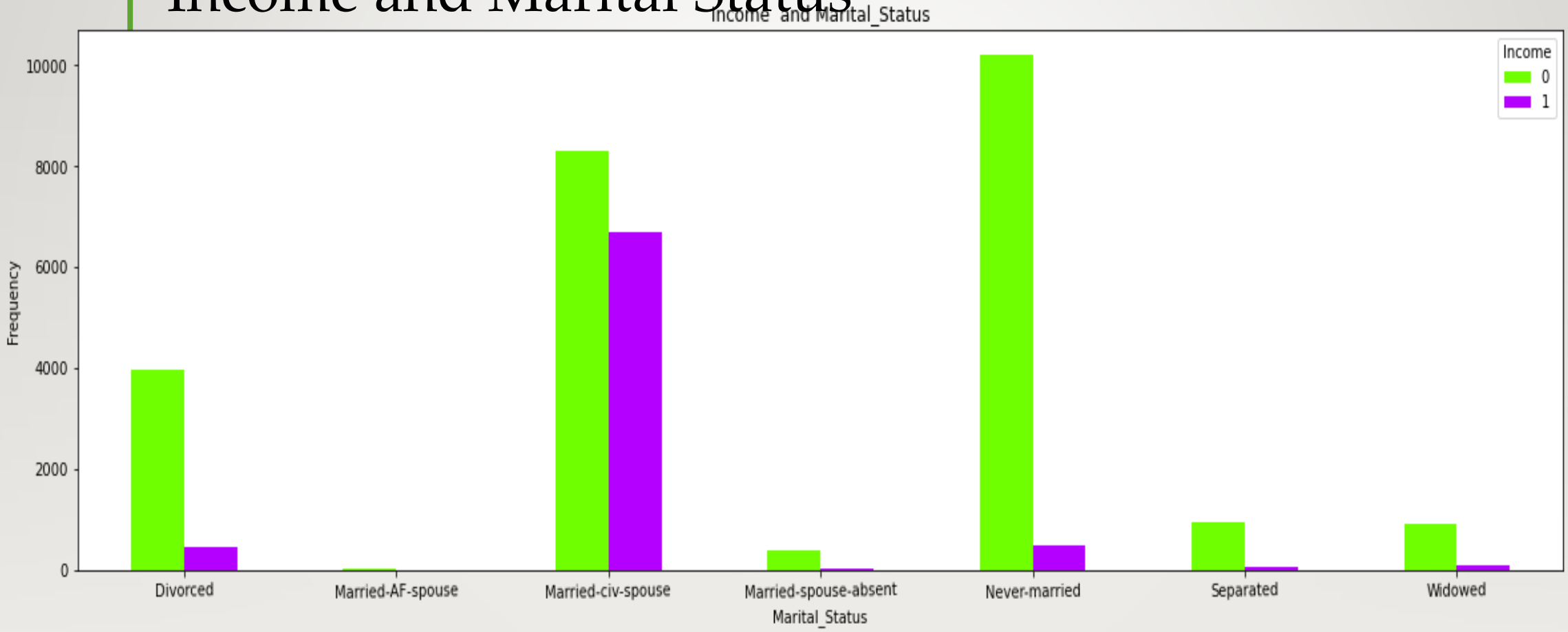
Comparing Age with Income



Comparing Sex with Income



Income and Marital Status



Pre-Processing and Modeling

	Age	Workclass	Final_Weight	Education	Education_Num	Marital_Status	Occupation	Relationship	Race	Sex	Capital_gain	Capital_loss	Hours_per_week	Native_country	Income
0	90	0	77053	11	9	6	0	1	4	0	0	4356	40	39	0
1	82	4	132870	11	9	6	4	1	4	0	0	4356	18	39	0
2	66	0	186061	15	10	6	0	4	2	0	0	4356	40	39	0
3	54	4	140359	5	4	0	7	4	4	0	0	3900	40	39	0
4	41	4	264663	15	10	5	10	3	4	0	0	3900	40	39	0

Qualitative variables in the dataset were encoded into numerical values

Multicollinearity Check

	VIF	Features
0	8.521265	Age
1	8.475260	Workclass
2	4.031573	Final_Weight
3	9.210325	Education
4	18.379729	Education_Num
5	3.976179	Marital_Status
6	3.710596	Occupation
7	2.612286	Relationship
8	17.578379	Race
9	4.441675	Sex
10	1.044798	Capital_gain
11	1.061817	Capital_loss
12	12.147293	Hours_per_week
13	19.726726	Native_country

Features with VIF(Variance inflation factor) greater than 10 will be dropped

Data for Modelling

	Age	Workclass	Final_Weight	Education	Marital_Status	Occupation	Relationship	Sex	Capital_gain	Capital_loss	Income
0	90	0	77053	11	6	0	1	0	0	4356	0
1	82	4	132870	11	6	4	1	0	0	4356	0
2	66	0	186061	15	6	0	4	0	0	4356	0
3	54	4	140359	5	0	7	4	0	0	3900	0
4	41	4	264663	15	5	10	3	0	0	3900	0

This is the final data which is spilt into test and train to be used for modelling

Modelling

The Model used in this Project are

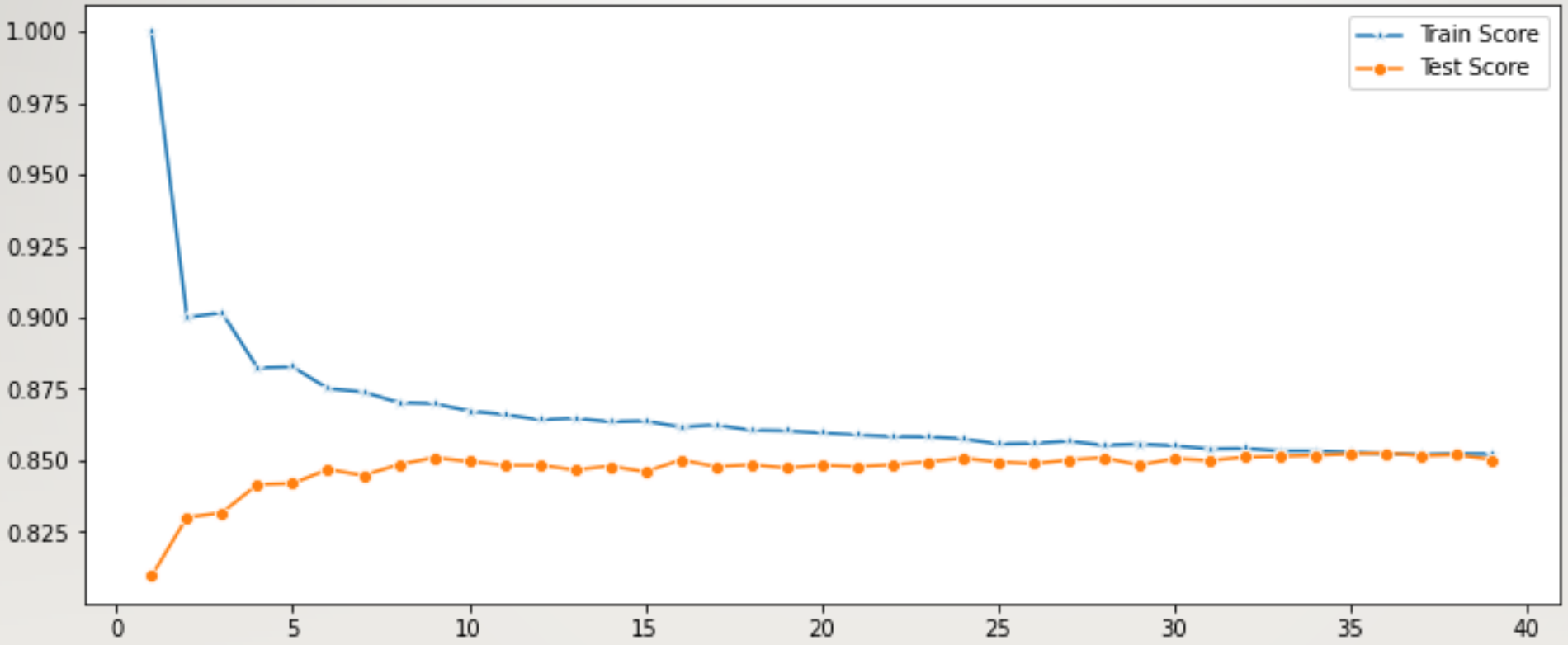
- Logistic Regression
- K-Nearest Neighbors Algorithm (KNN)

Logistic Regression Model (LRM)

The Training Accuracy is: 0.801478979321678
The Testing Accuracy is: 0.8064410121590536

	precision	recall	f1-score	support
0.0	0.81	0.97	0.88	17004
1.0	0.68	0.21	0.32	4903
accuracy			0.80	21907
macro avg	0.75	0.59	0.60	21907
weighted avg	0.78	0.80	0.76	21907

K-Nearest Neighbors Algorithm



k=36 has the highest score for the testing data

K-Nearest Neighbors Algorithm

WITH K=36

```
[[4444  282]
 [ 615  745]]
```

	precision	recall	f1-score	support
0.0	0.88	0.94	0.91	4726
1.0	0.73	0.55	0.62	1360
accuracy			0.85	6086
macro avg	0.80	0.74	0.77	6086
weighted avg	0.84	0.85	0.84	6086

End

Thank You Very Much