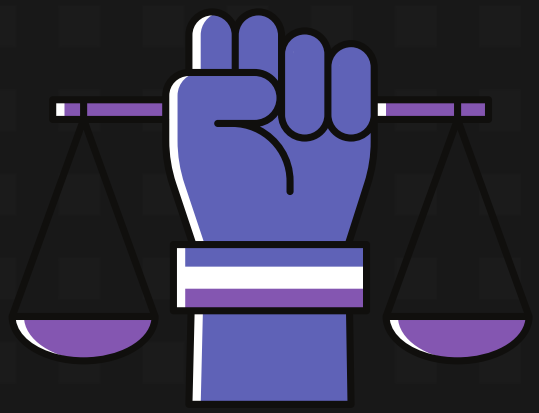
01

GENDER WAGE GAP AND SALARY PREDICTION

BERRA KARAYEL



PROBLEM STATEMENT

In today's society, gender pay gap remains a significant issue across various industries and professions worldwide. Understanding the factors influencing salary discrepancies between genders is crucial for promoting fairness and equality in the workplace. This project aims to analyze the Gender Pay Gap using the 2018 Kaggle Machine Learning & Data Science Survey dataset, focusing on factors such as age, education, profession, industry, and experience.

BERRA KARAYEL

DATA SET

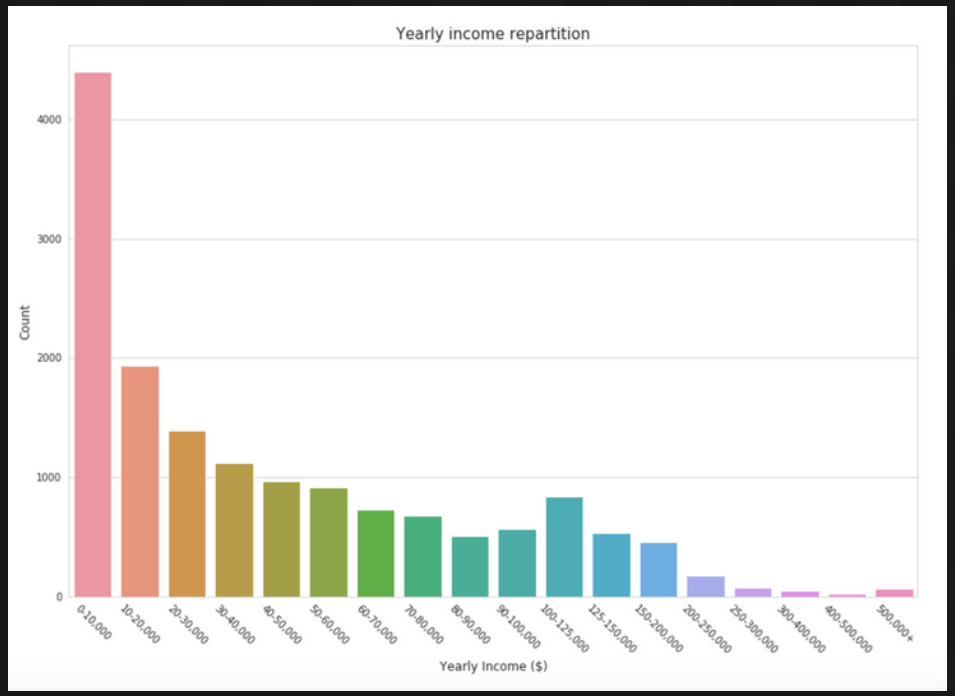
2018 Kaggle Machine Learning & Data Science Survey Data

In this project, the questions that are used:

- Gender
- Age
- Nationality
- Education
- Major
- Profession
- Industry of profession
- Ancienety
- The annual income

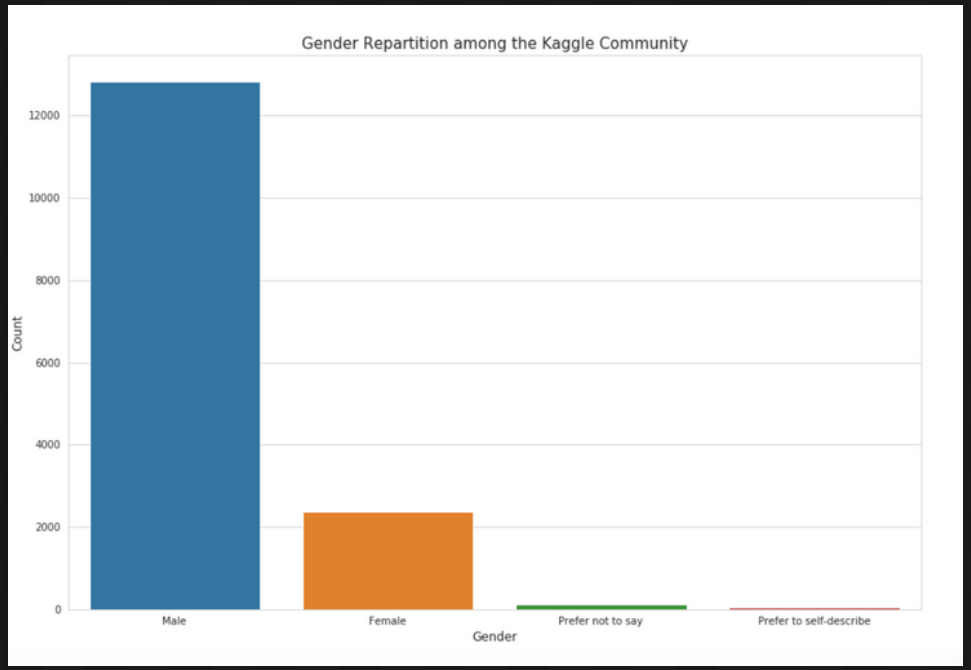
METHODS

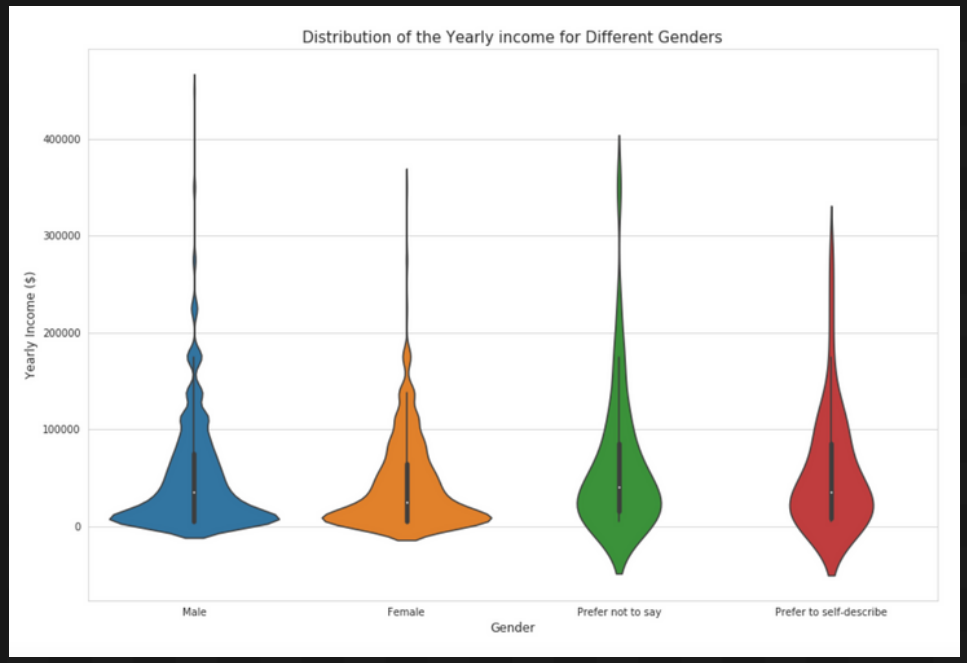
- Target visualization
- Exploratory Data Analysis & Gender Wage Gap Analysis
- Machine Learning Model Design

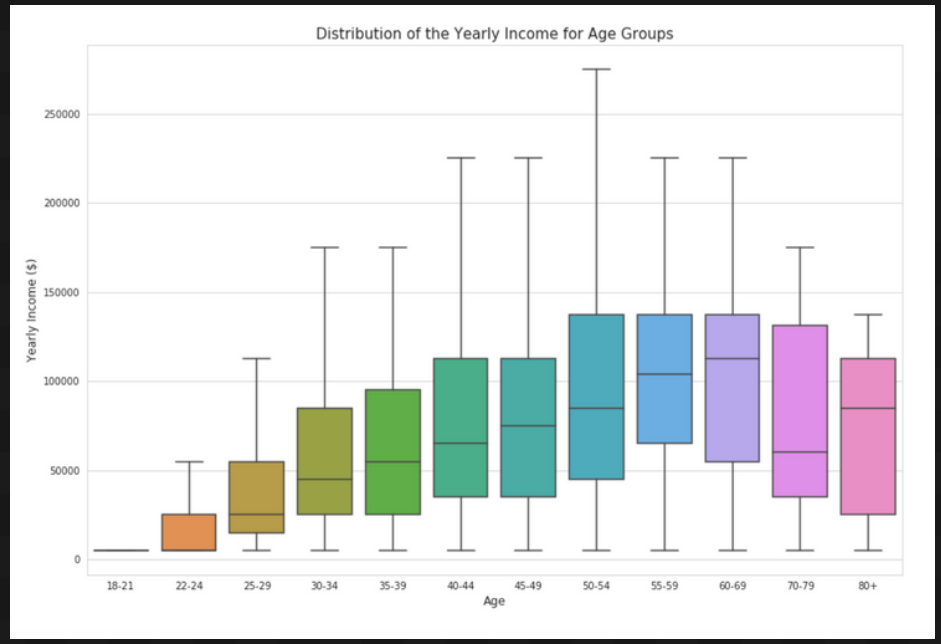


BERRA KARAYEL

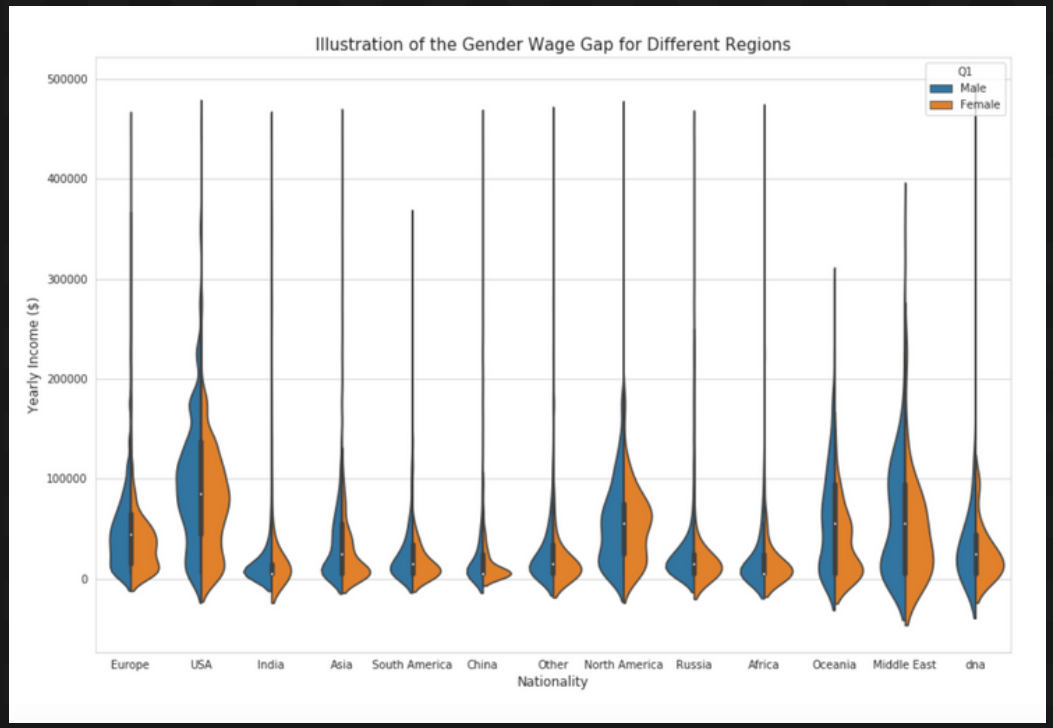
EXPLORATORY DATA ANALYSIS

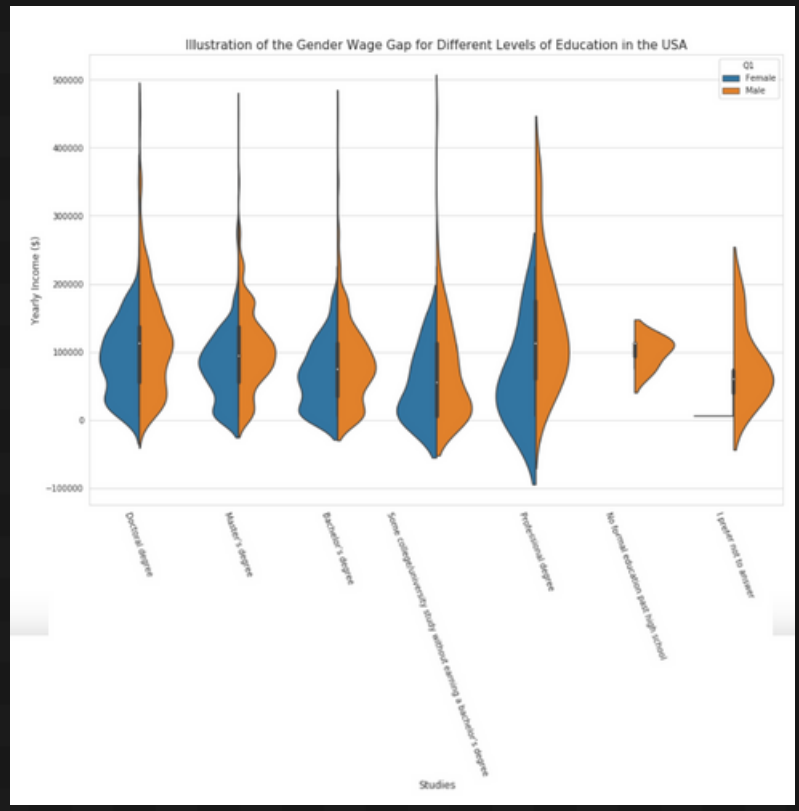



Distribution of the Yearly income for Different Genders
Yearly Income ($)
400000
300000
200000
100000
0
Male
Female
Prefer not to say
Prefer to self-describe
Gender

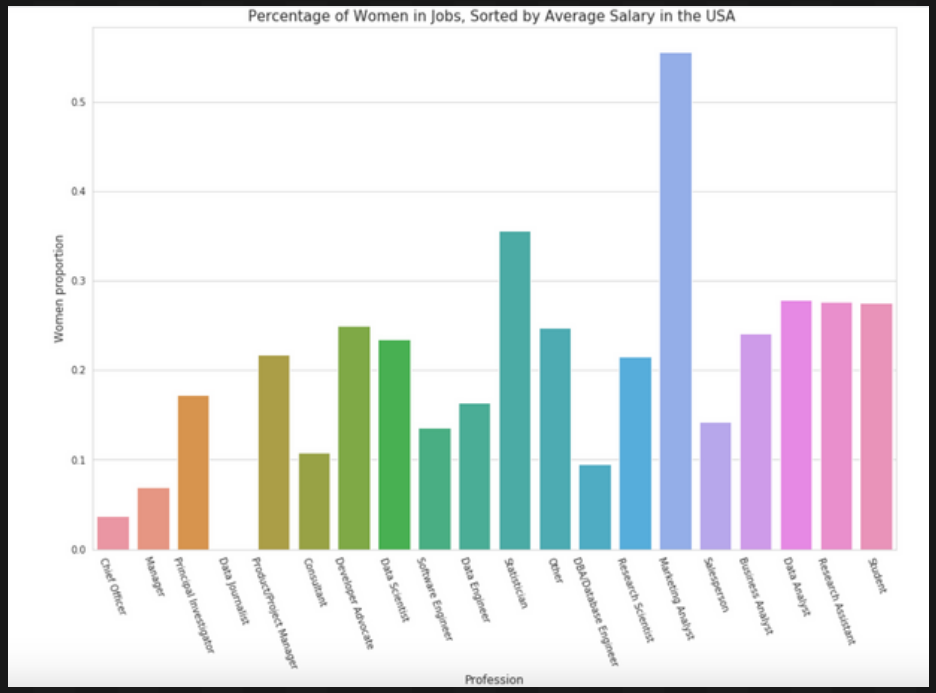

Distribution of the Yearly Income for Age Groups
Yearly Income ($)
250000
200000
150000
100000
50000
0
18-21
22-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-69
70-79
80+
Age

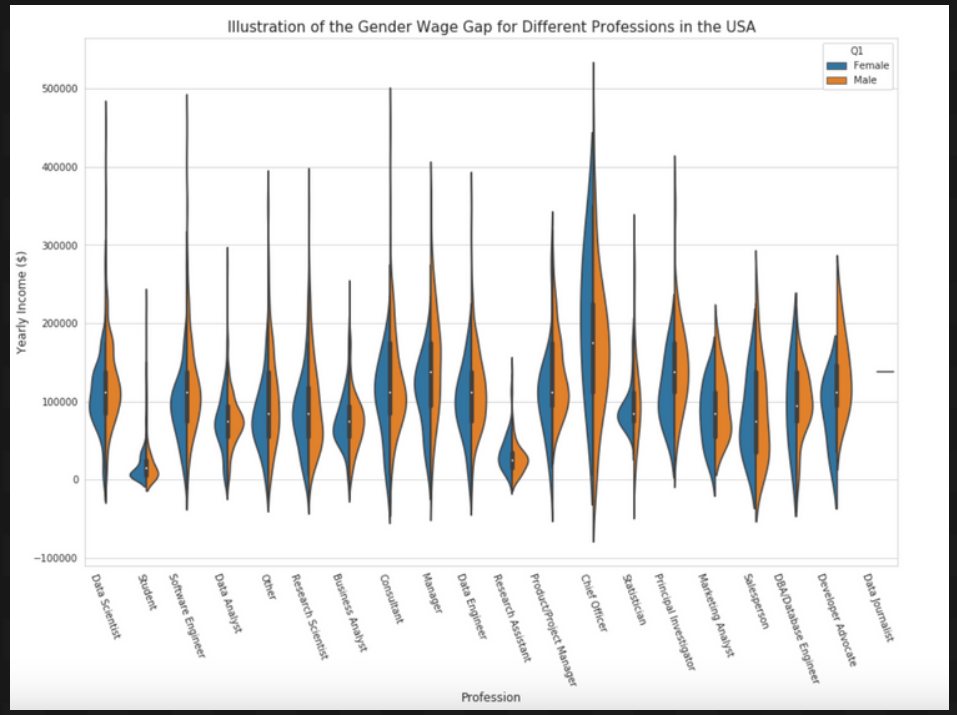
BERRA KARAYEL



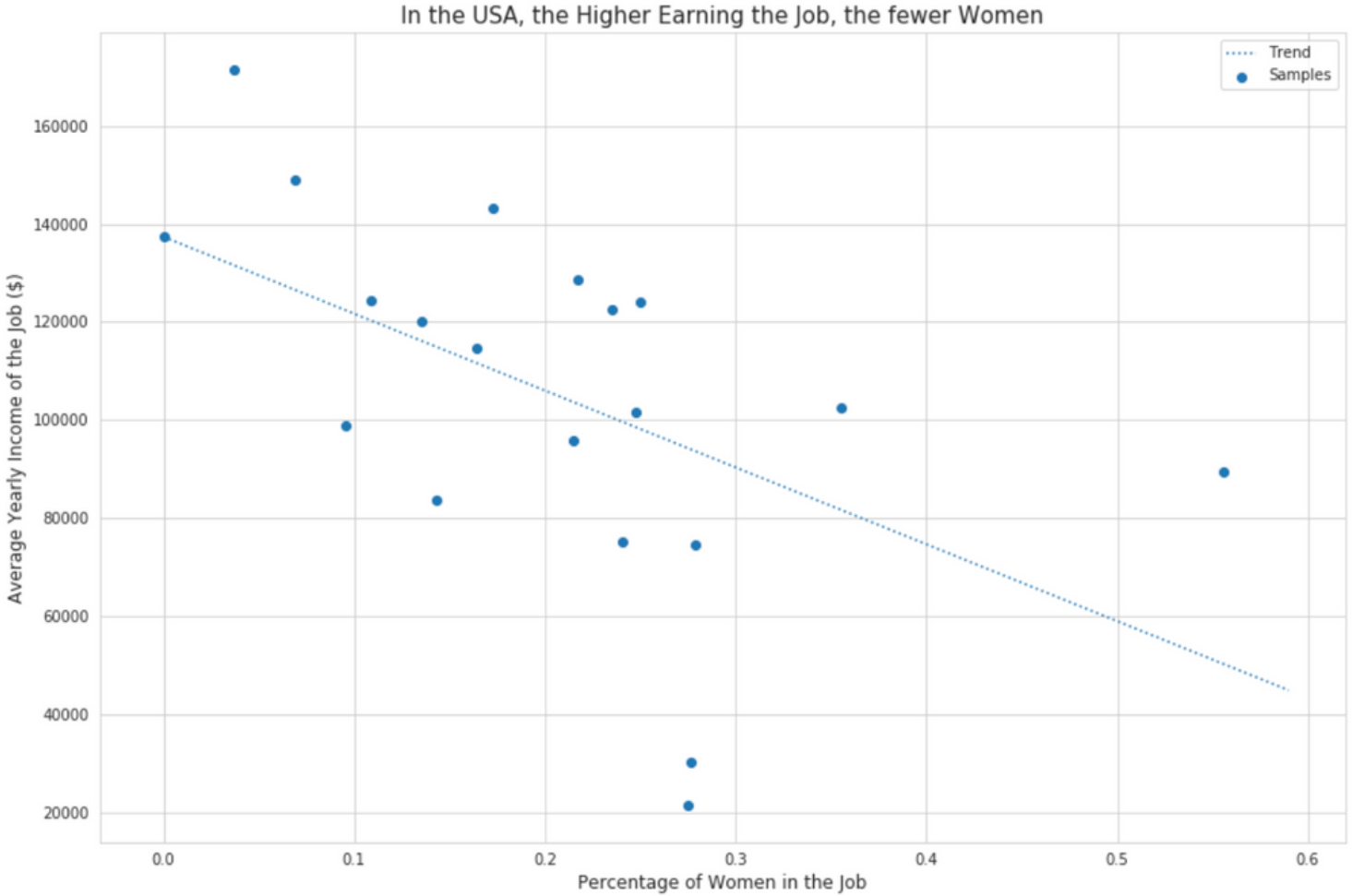


BERRA KARAYEL

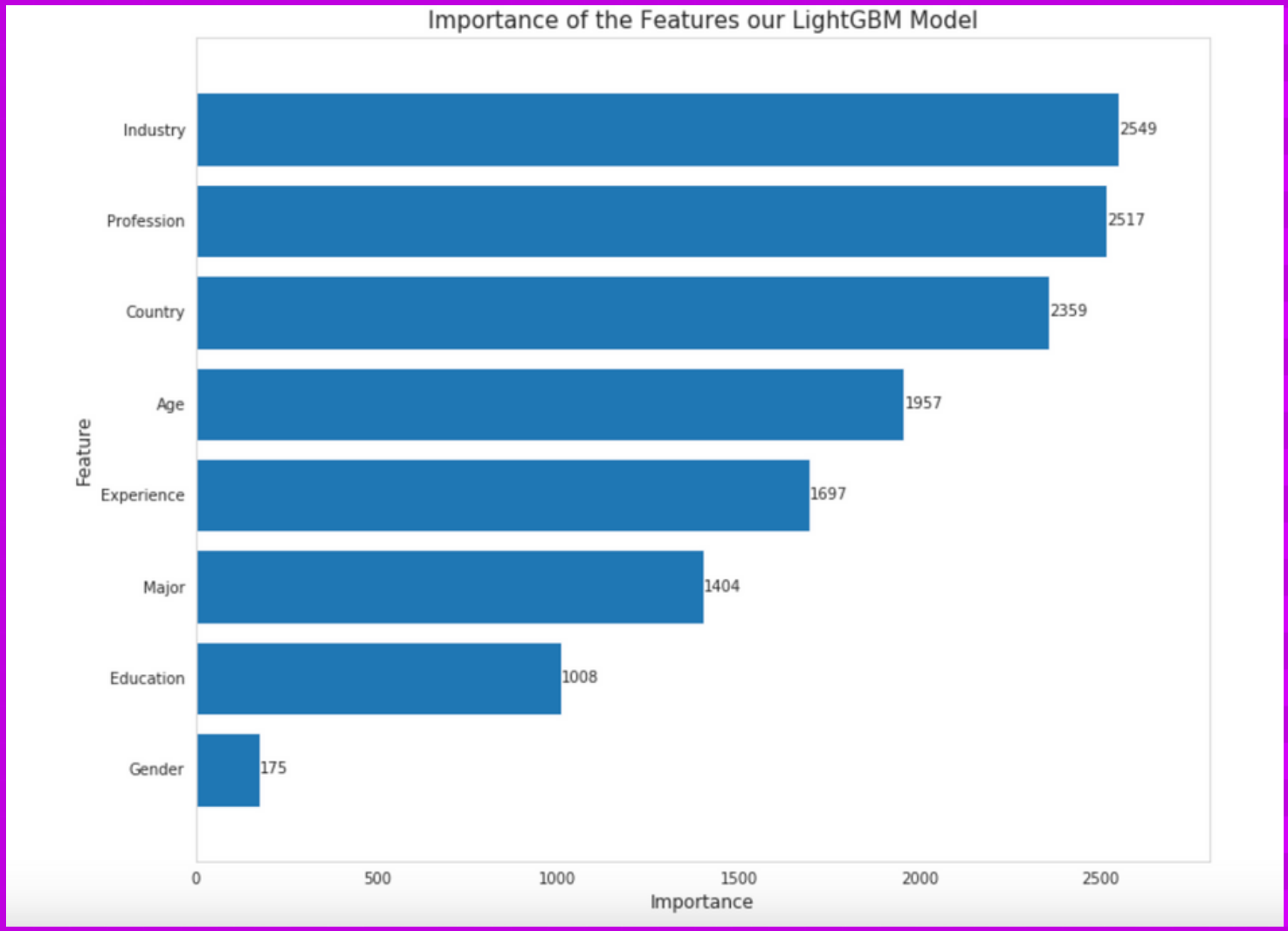


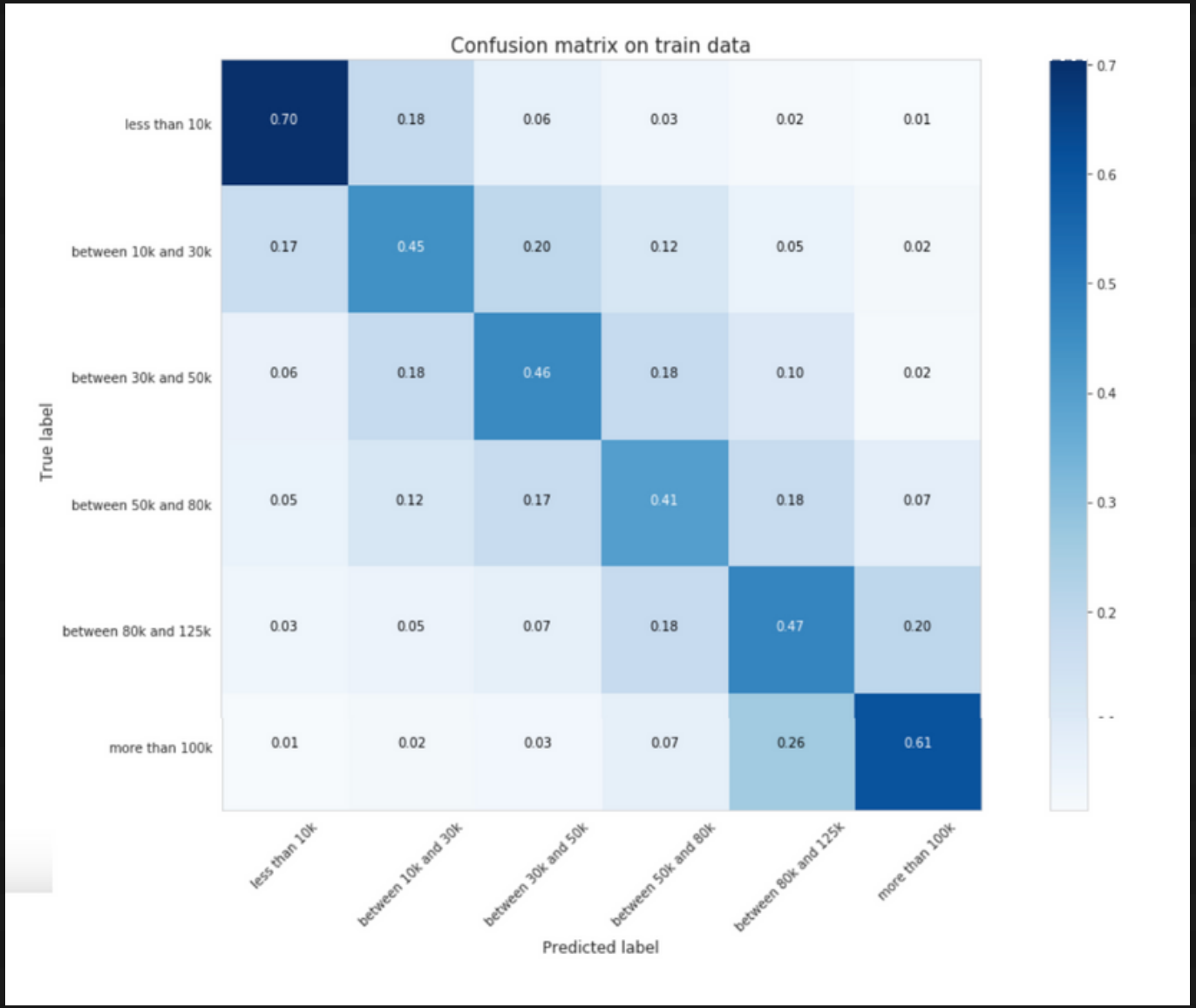


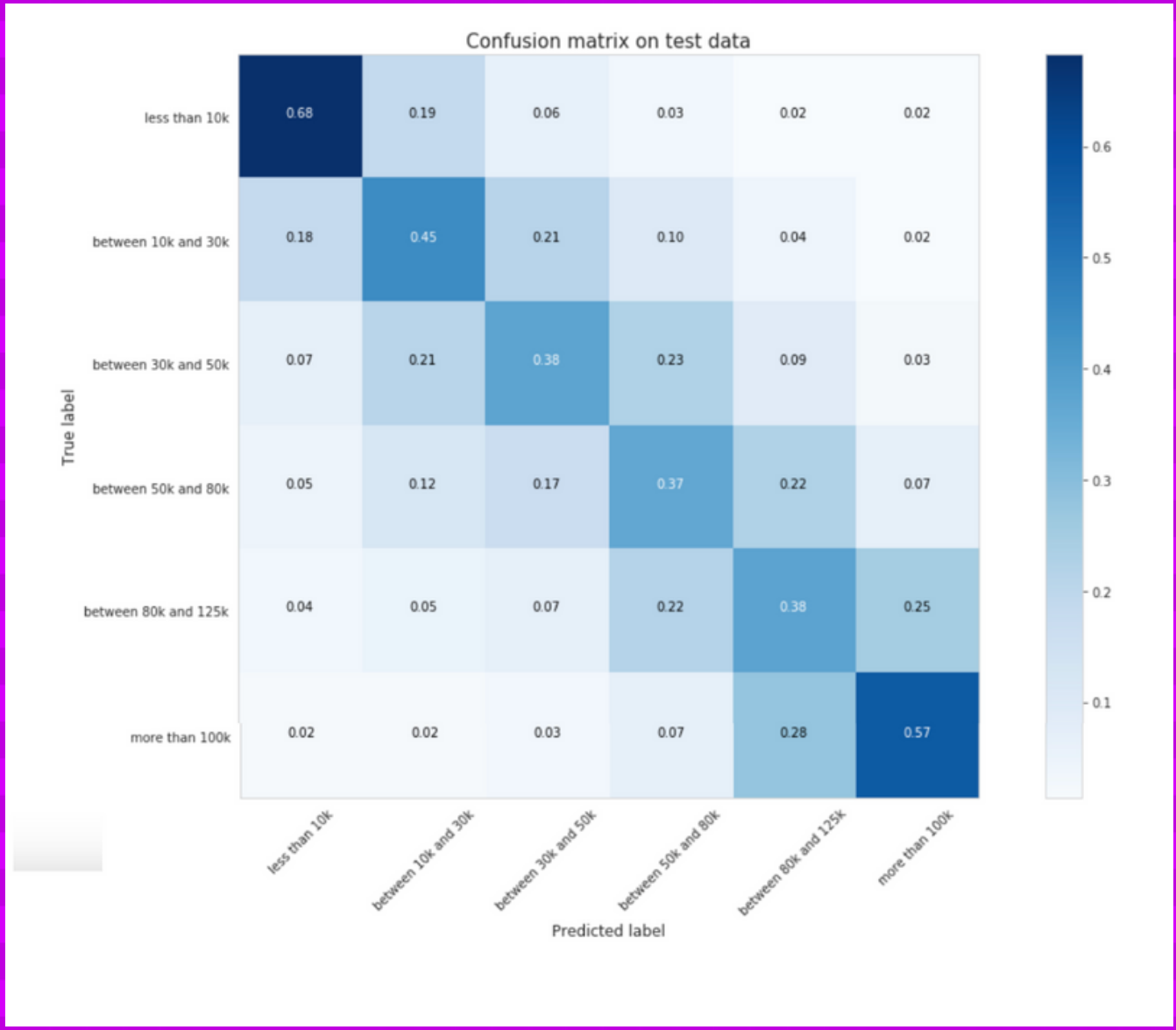
08


In the USA, the Higher Earning the Job, the fewer Women
Trend
Samples
Average Yearly Income of the Job ($)
160000
140000
120000
100000
80000
60000
40000
20000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Percentage of Women in the Job

GRADIENT BOOSTING WITH LIGTHGBM




Confusion matrix on train data
less than 10k
between 10k and 30k
between 30k and 50k
between 50k and 80k
between 80k and 125k
more than 100k
True label
0.70 0.18 0.06 0.03 0.02 0.01
0.17 0.45 0.20 0.12 0.05 0.02
0.06 0.18 0.46 0.18 0.10 0.02
0.05 0.12 0.17 0.41 0.18 0.07
0.03 0.05 0.07 0.18 0.47 0.20
0.01 0.02 0.03 0.07 0.26 0.61
less than 10k
between 10k and 30k
between 30k and 50k
between 50k and 80k
between 80k and 125k
more than 100k
Predicted label
0.7
0.6
0.5
0.4
0.3
0.2


Confusion matrix on test data
less than 10k
between 10k and 30k
between 30k and 50k
between 50k and 80k
between 80k and 125k
more than 100k
True label
0.68 0.19 0.06 0.03 0.02 0.02
0.18 0.45 0.21 0.10 0.04 0.02
0.07 0.21 0.38 0.23 0.09 0.03
0.05 0.12 0.17 0.37 0.22 0.07
0.04 0.05 0.07 0.22 0.38 0.25
0.02 0.02 0.03 0.07 0.28 0.57
less than 10k
between 10k and 30k
between 30k and 50k
between 50k and 80k
between 80k and 125k
more than 100k
Predicted label
0.6
0.5
0.4
0.3
0.2
0.1

THANK YOU!

Berra Karayel