Machine Learning final Project

ME

Eyal Michaeli

Economics & Data analytics graduate

BIG DATA

What are we trying to predict?

We are trying to predict the **Default of Credit Card Clients - Classification**

Our Data set

Demographic Data

Education, age, sex, marriage status

Limit Balance

Amount of given credit in NT dollars

Repayment status – 6 monts back

Wheather a client has payed dully, and if not, by how many months back?

bill **statement** – 6 months back

Amount of bill statement

Previous payment – 6 months back

Amount of previous payment in each of the last 6 months

Default

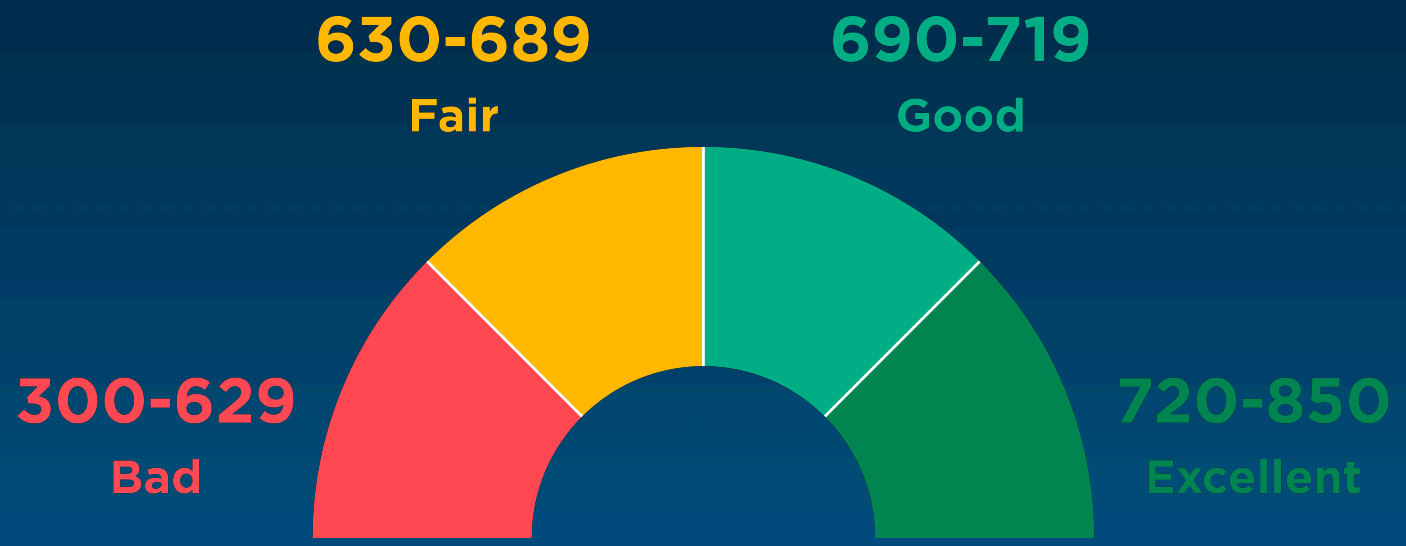
If a client has defaulted

Motivation and applications for solving the problem

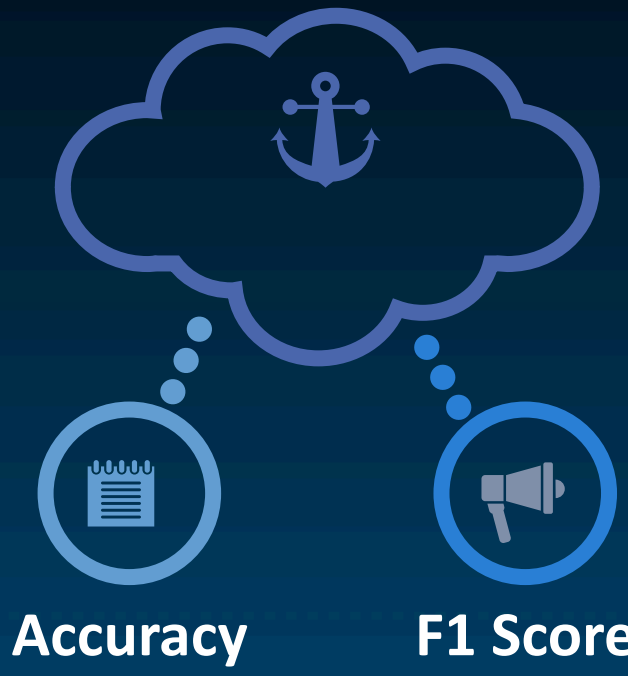
As a basis for credit card Rating

Give Higher interest rate for ‘dangerous’ clients

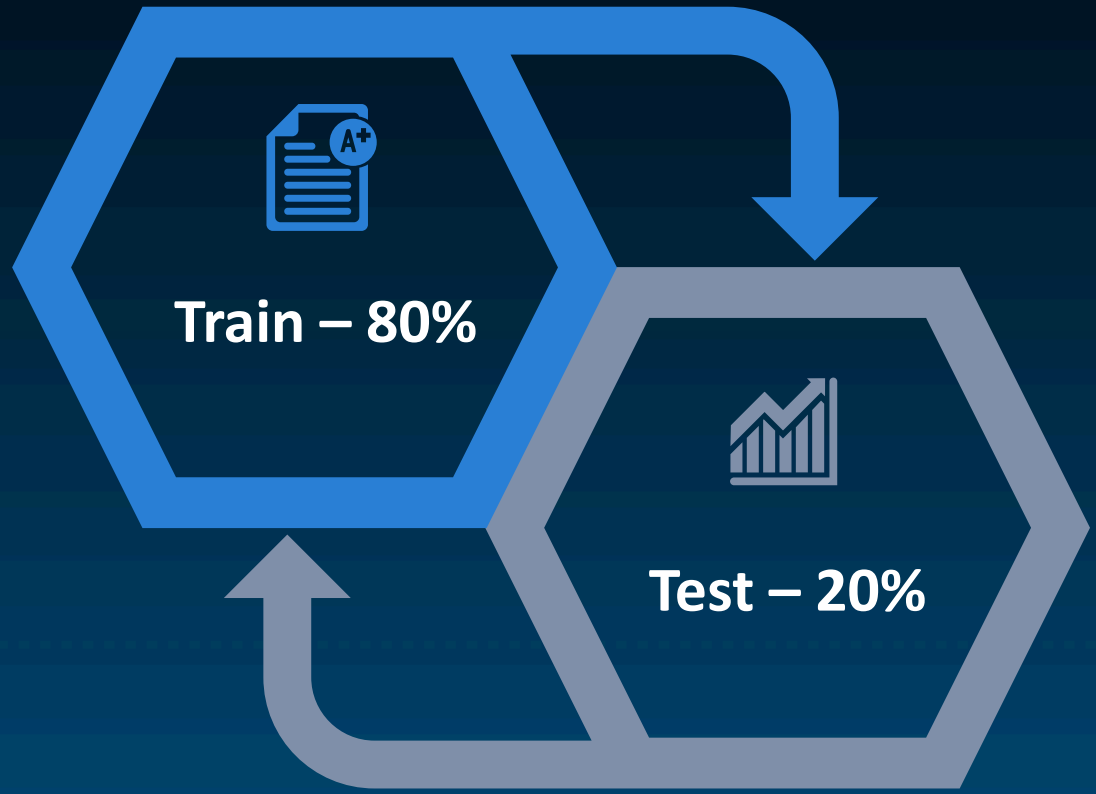
Don’t give or give less credit



Evaluation



Train – test split



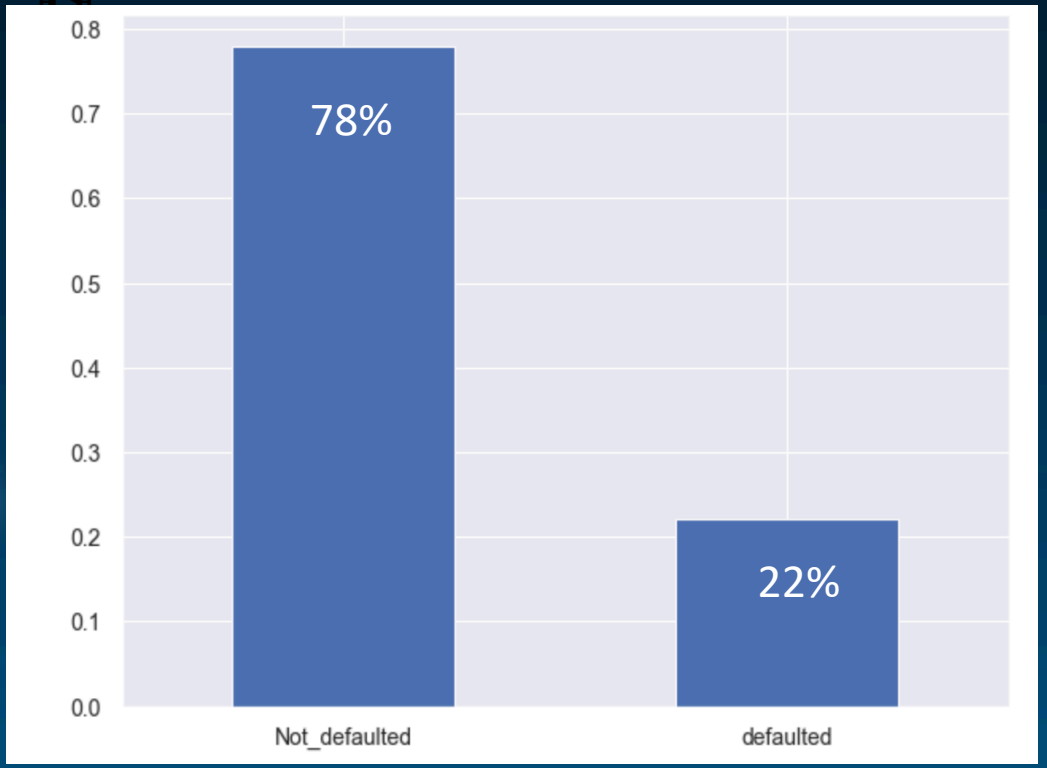
Data description

24 features

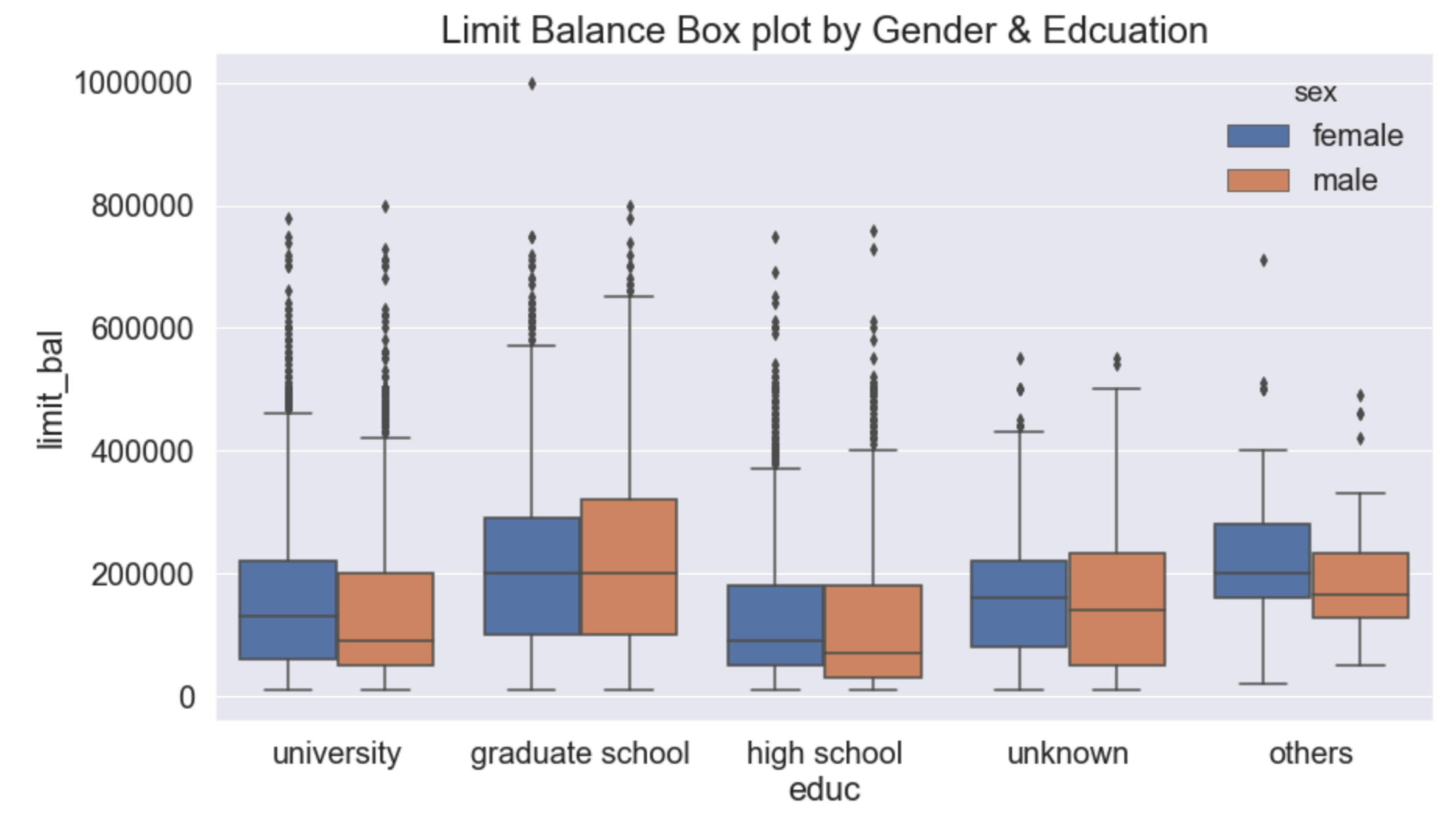
30K observations

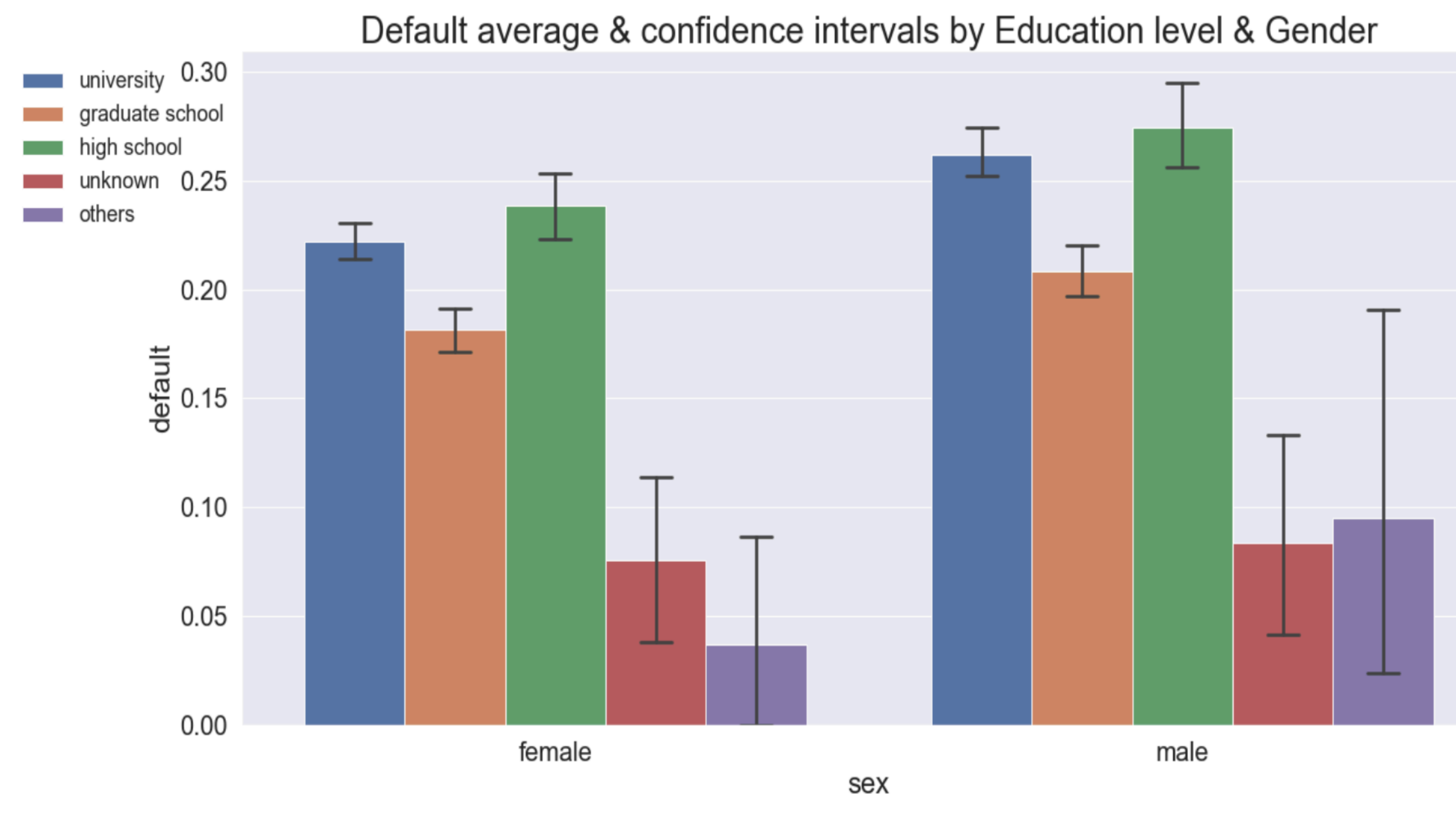
Small unbalance in labels:

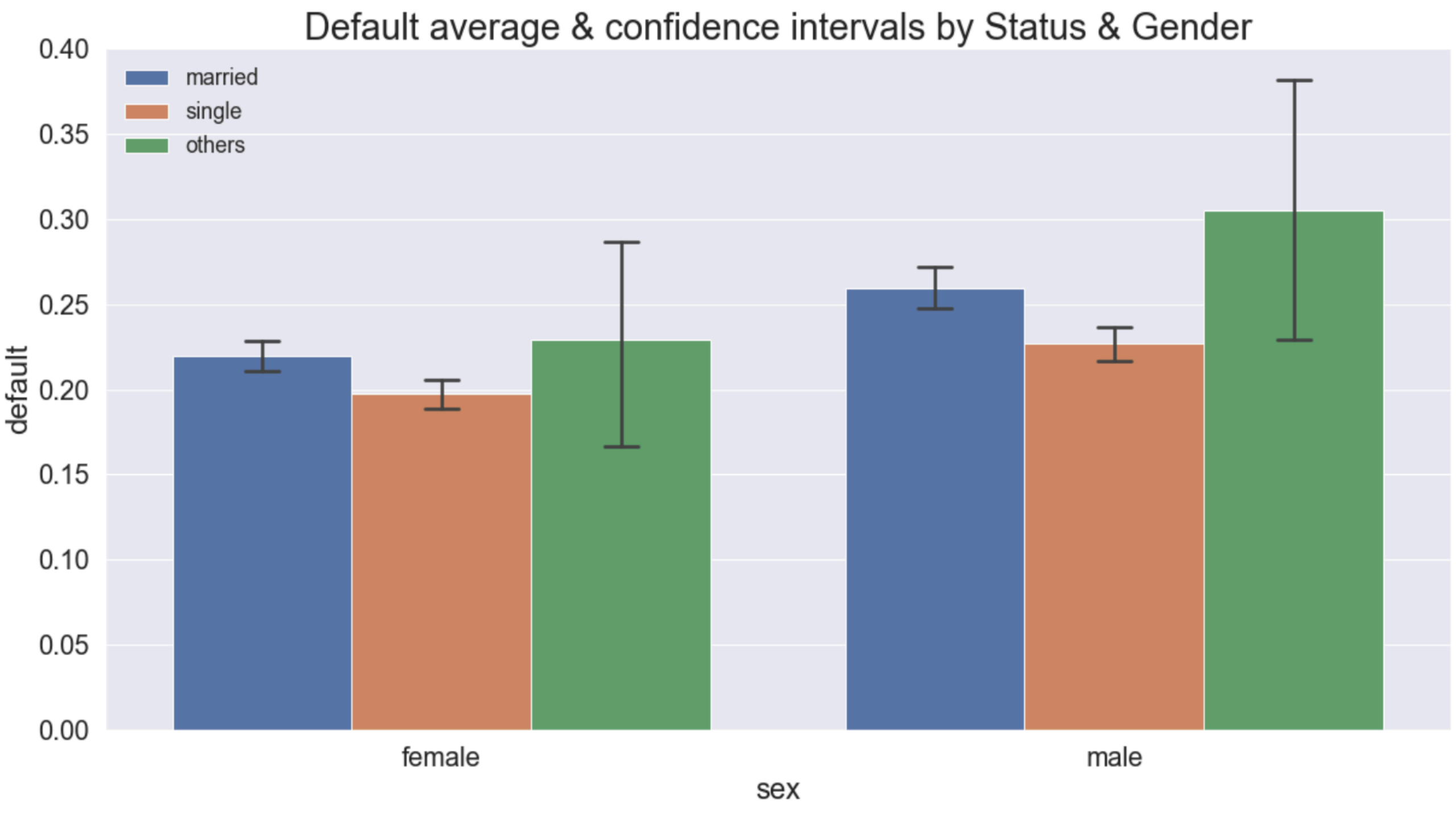
NA – only 6& NA values(less than 0.1%)

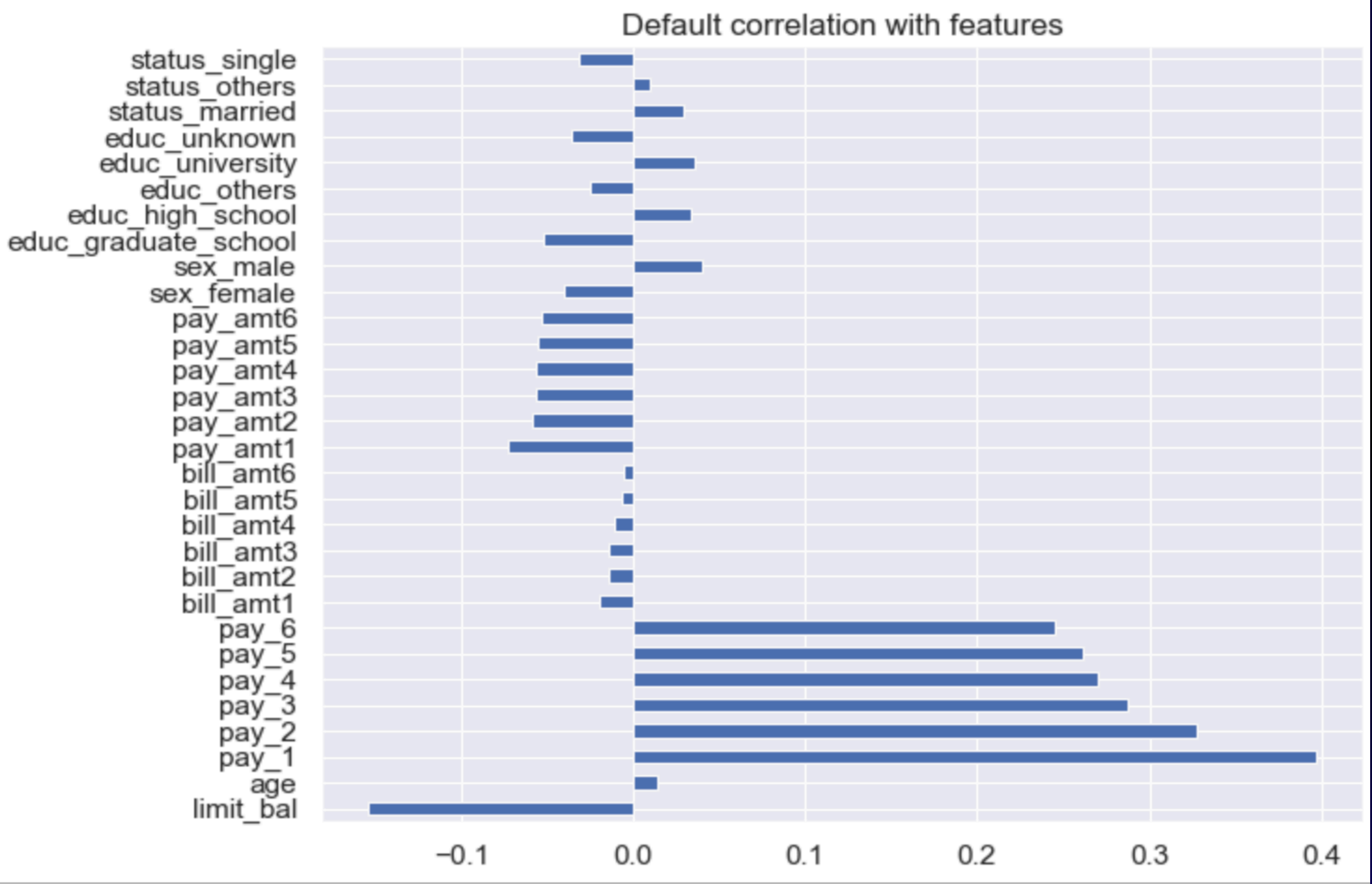


- Graphs describing various aspects of the data

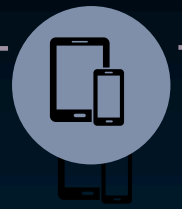

Limit Balance Box plot by Gender & Edcuation
limit_bal
1000000
800000
600000
400000
200000
0
sex
female
male
university
graduate school
high school
unknown
others
educ


Default average & confidence intervals by Education level & Gender
university
graduate school
high school
unknown
others
0.30
0.25
0.20
0.15
0.10
0.05
0.00
default
female
male
sex


Default average & confidence intervals by Status & Gender
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
default
married
single
others
female
male
sex


Default correlation with features
status_single
status_others
status_married
educ_unknown
educ_university
educ_others
educ_high_school
educ_graduate_school
sex_male
sex_female
pay_amt6
pay_amt5
pay_amt4
pay_amt3
pay_amt2
pay_amt1
bill_amt6
bill_amt5
bill_amt4
bill_amt3
bill_amt2
bill_amt1
pay_6
pay_5
pay_4
pay_3
pay_2
pay_1
age
limit_bal
−0.1
0.0
0.1
0.2
0.3
0.4

Data engineering



Did you add any features?

One-Hot-Encoding

Dropped rows with NA values

Removed ID column

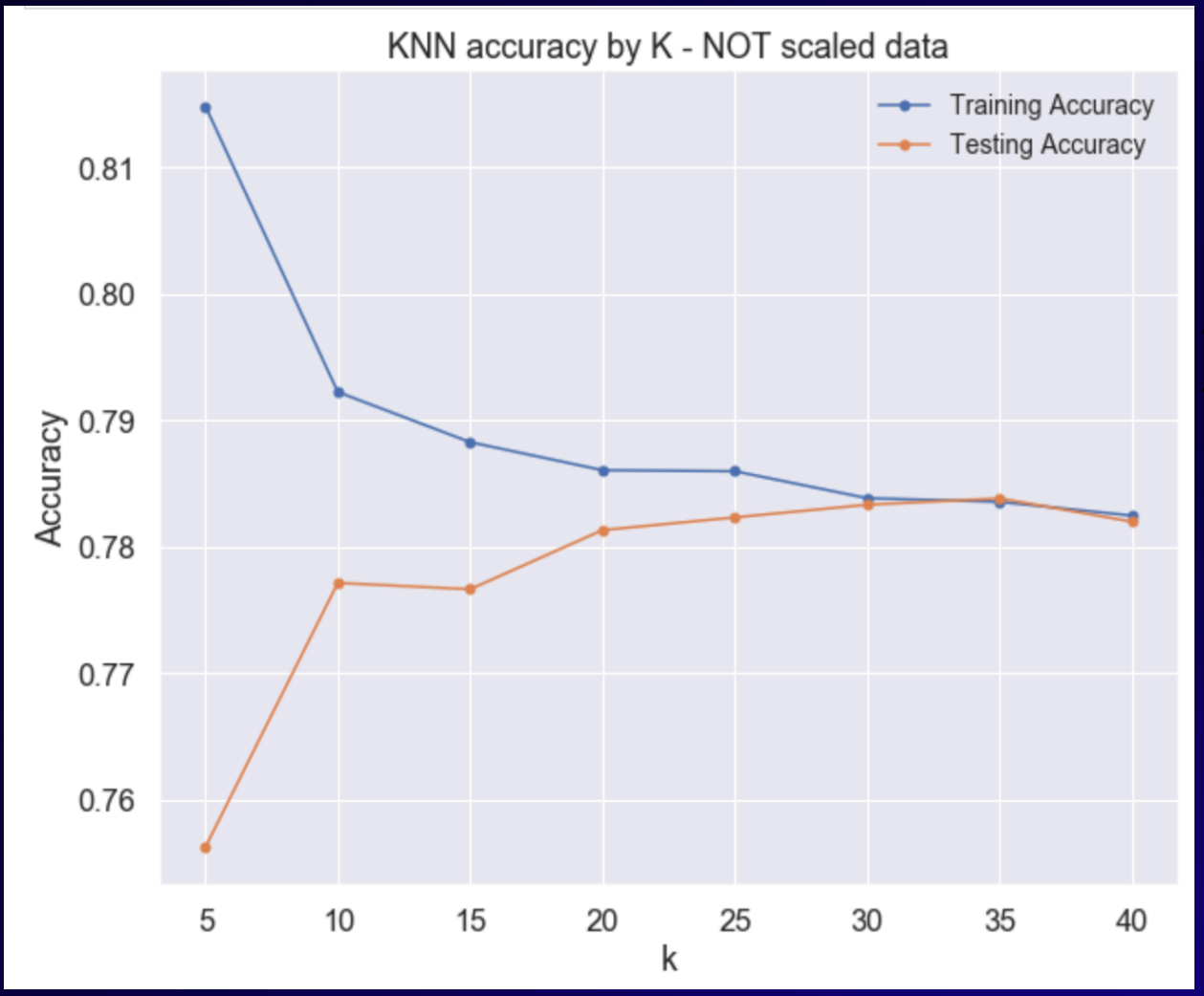
Benchmark algorithm

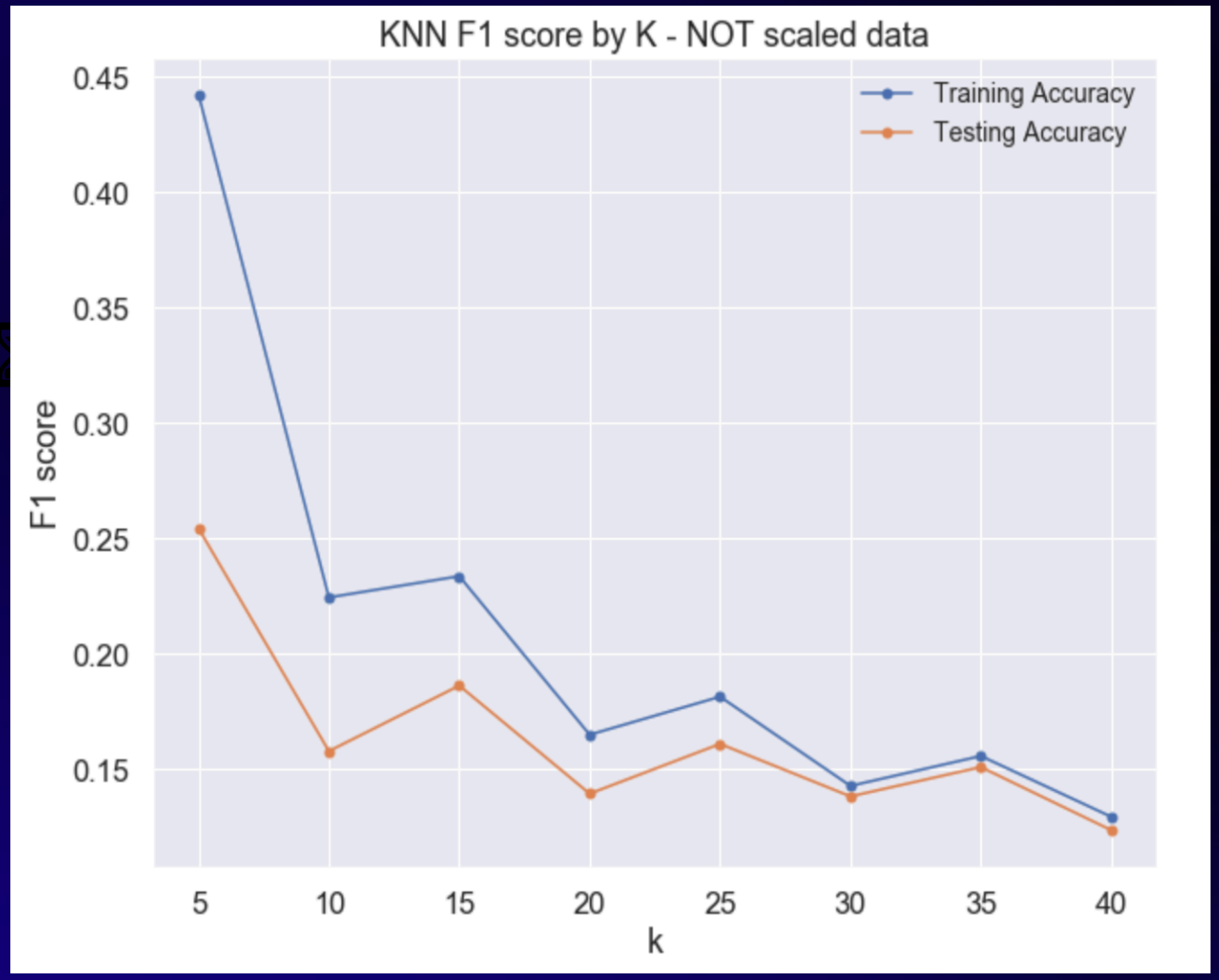
Prediction – always not default

Accuracy of 78%
F1 score of 0(of course).

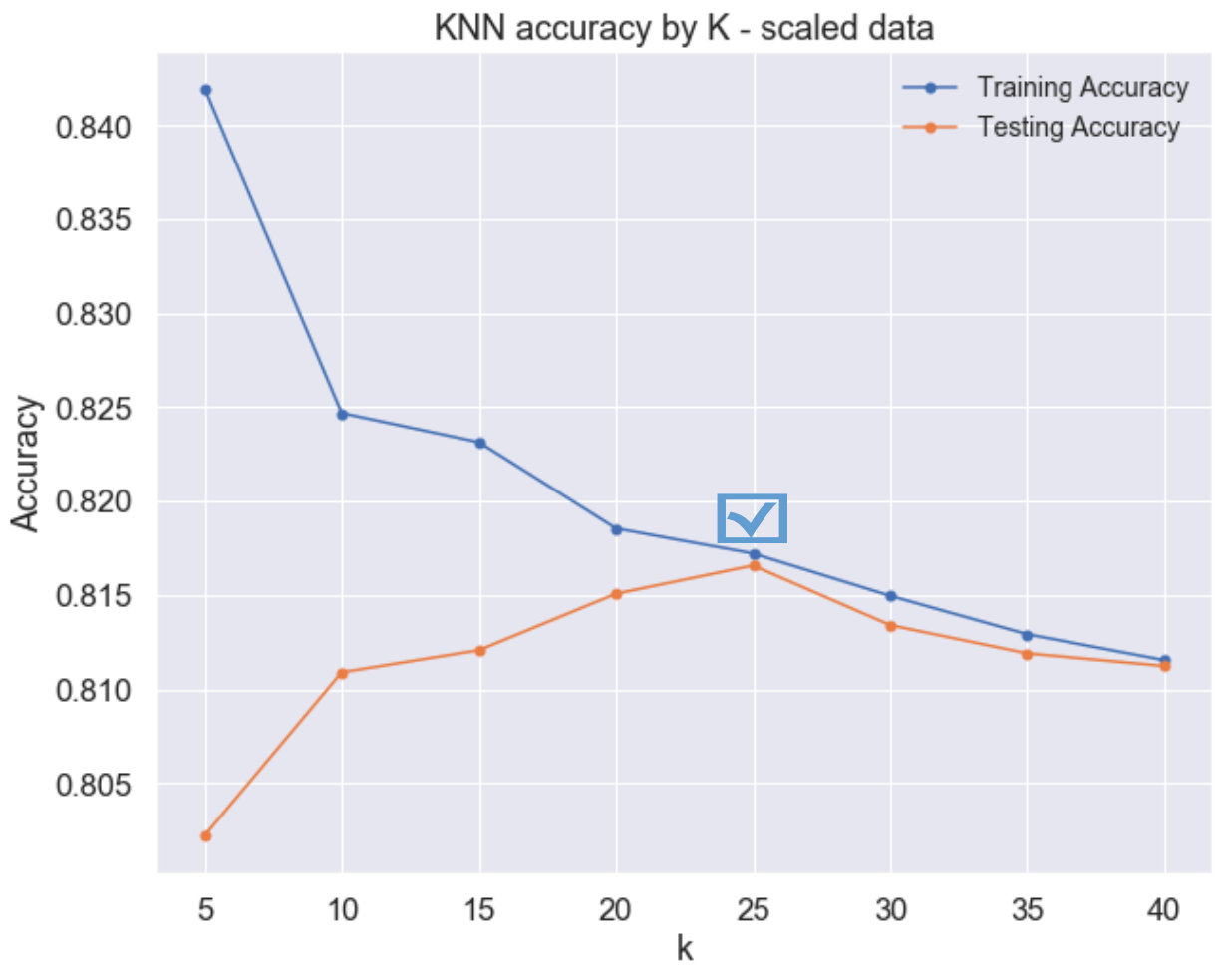
We are going to try to beat that!

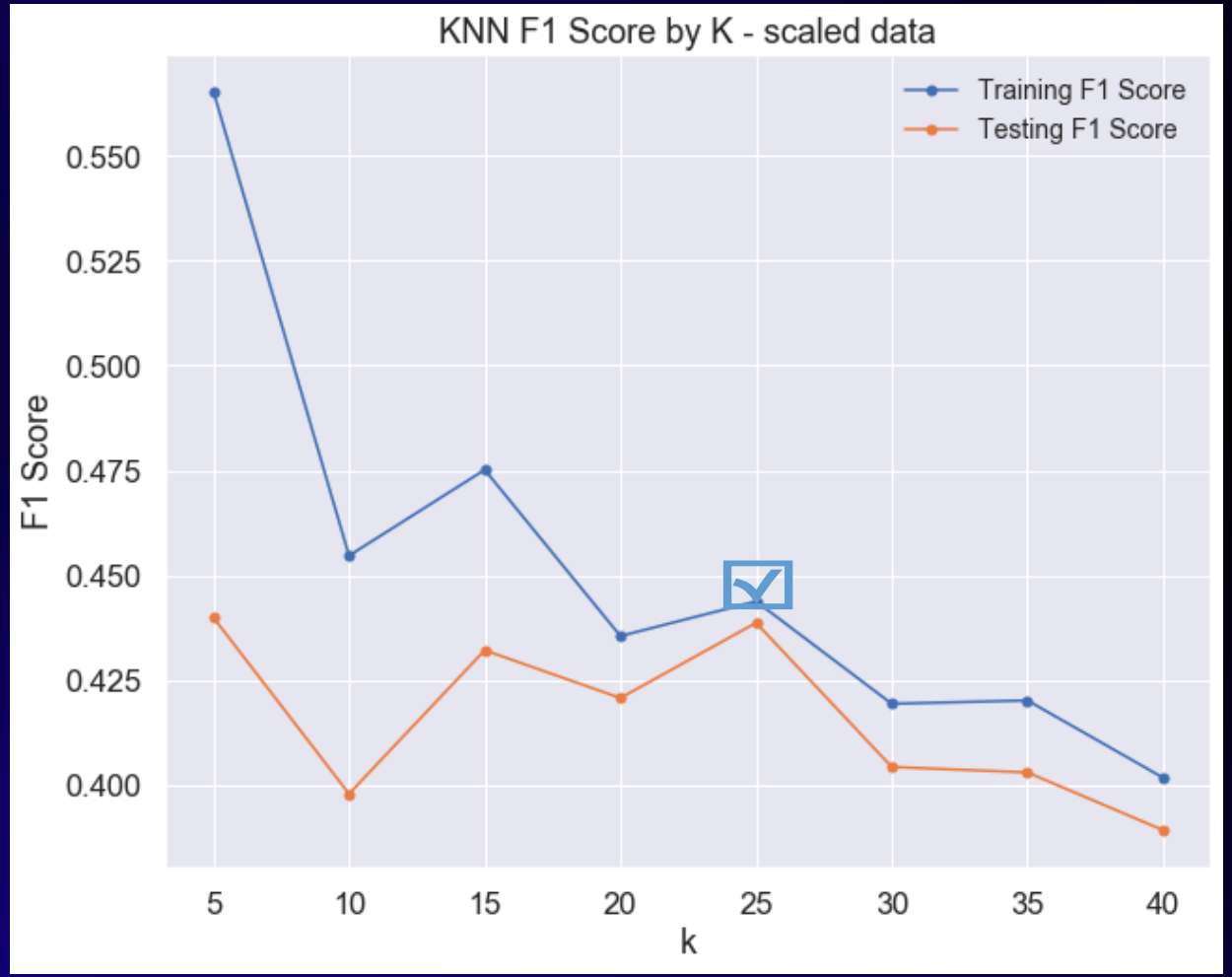
KNN – NOT scaled data



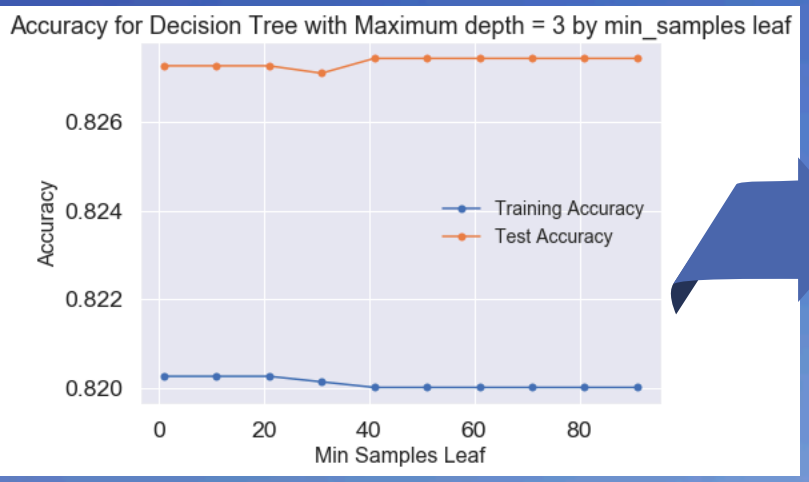


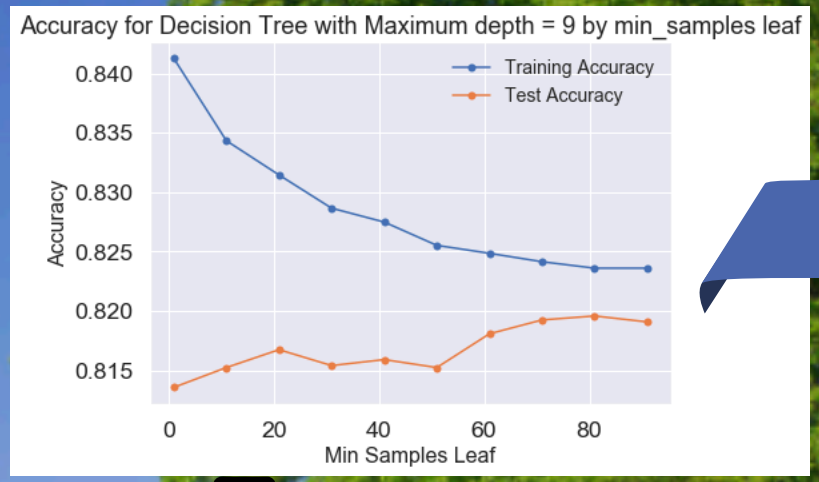
KNN –scaled data

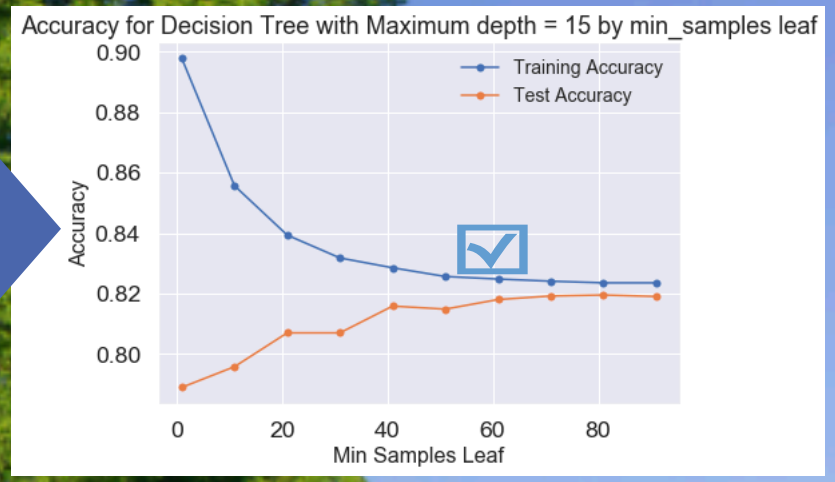


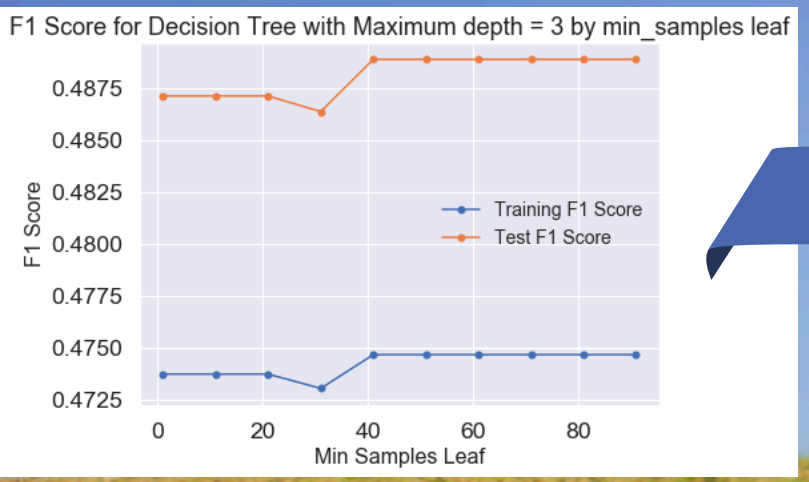


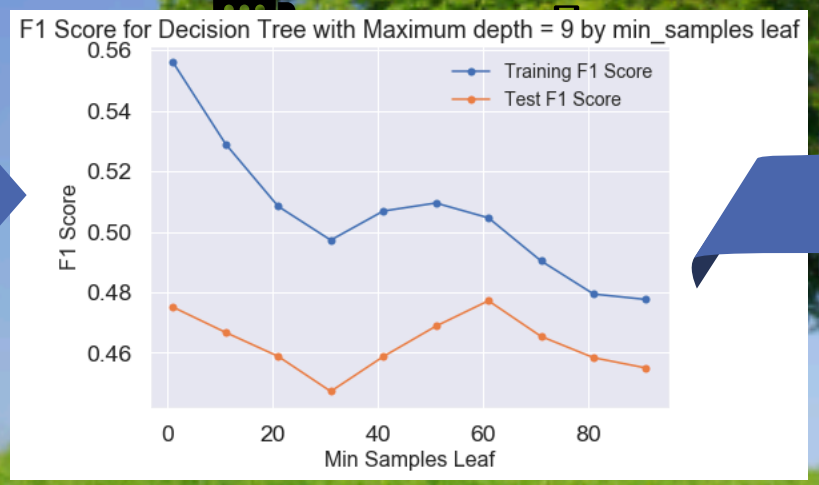
Decision tree

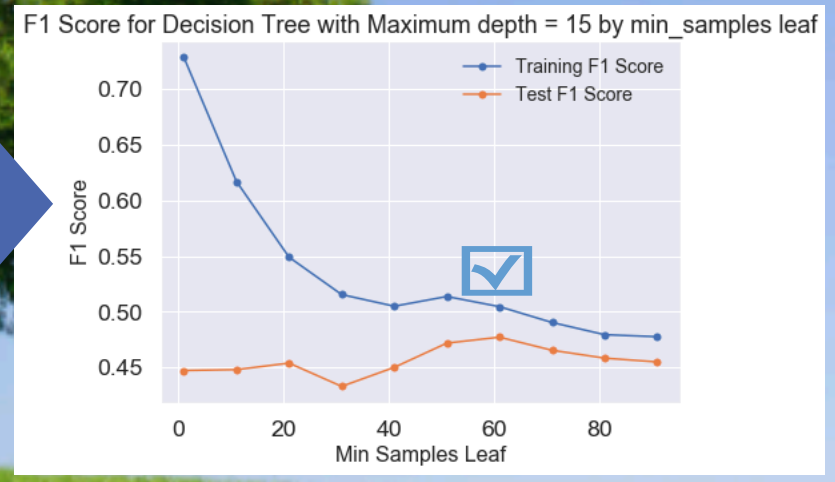






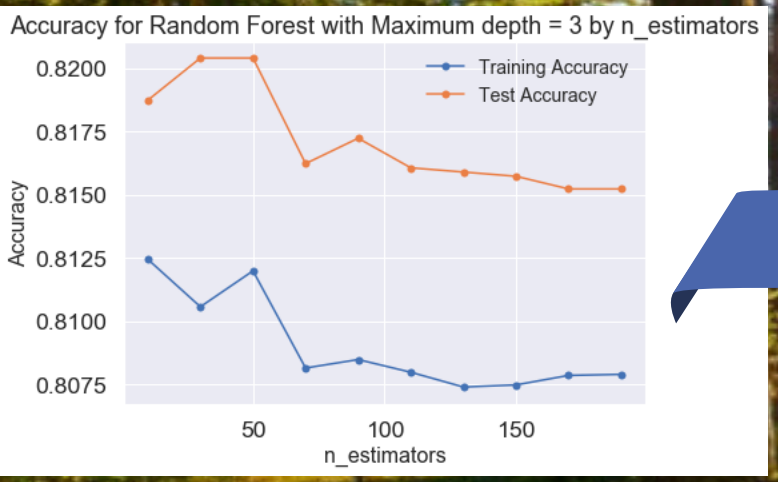


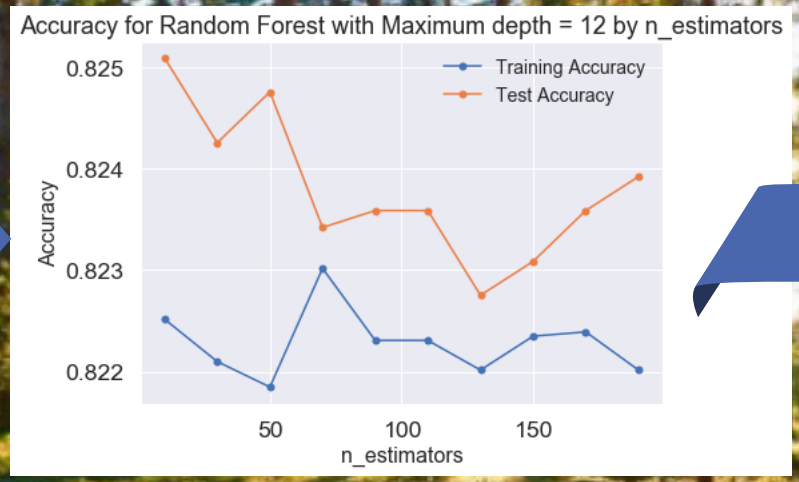


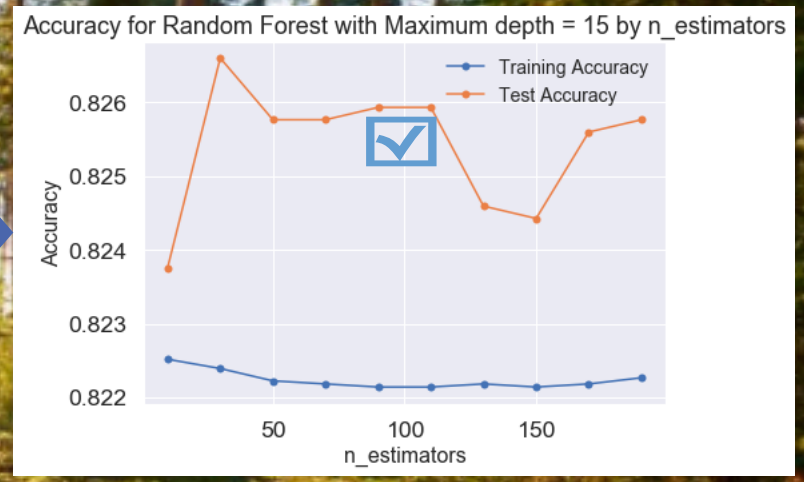


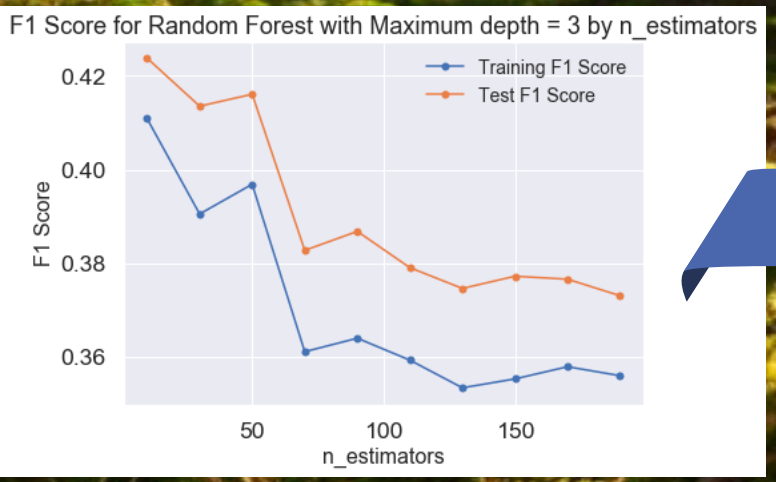
We chose max depth of 15 and Min samples leaf of 60

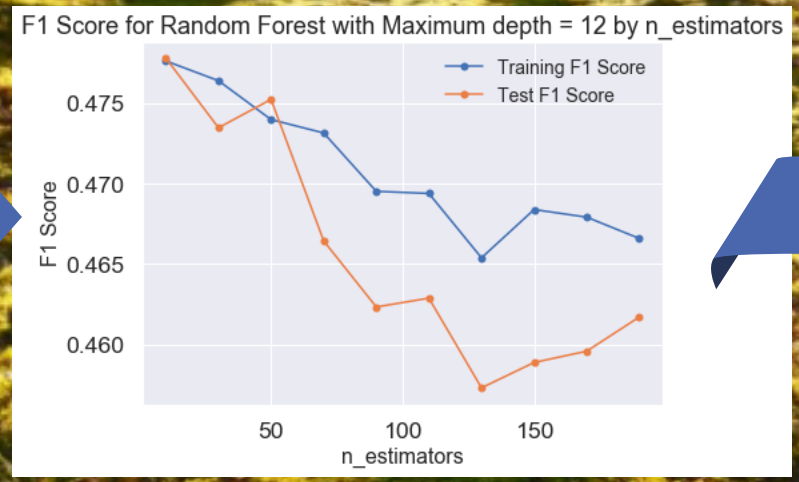
Random Forest

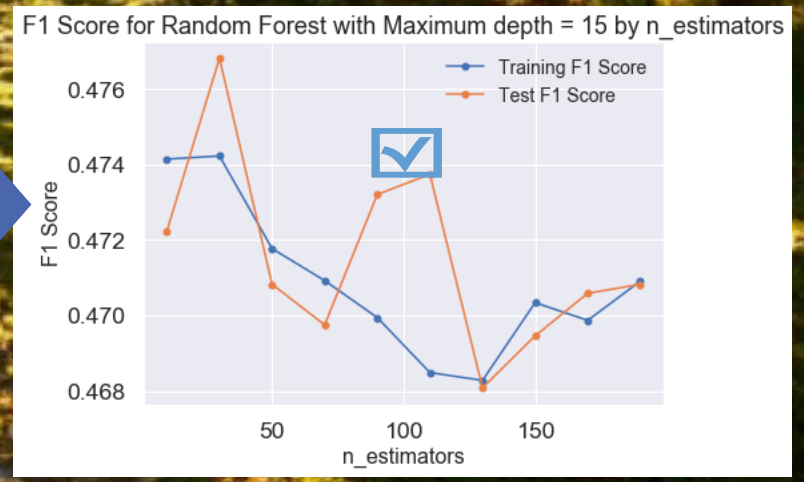






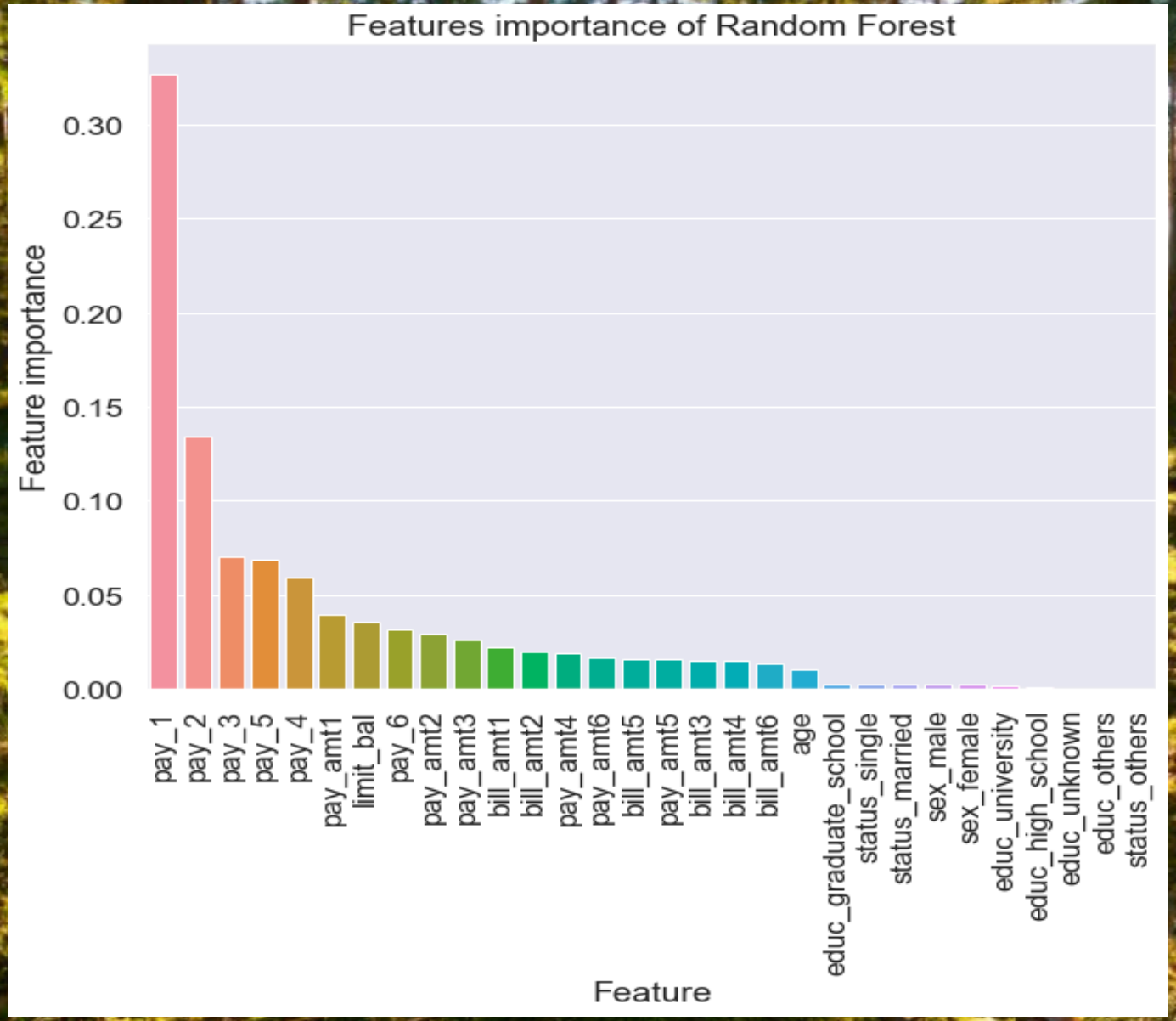




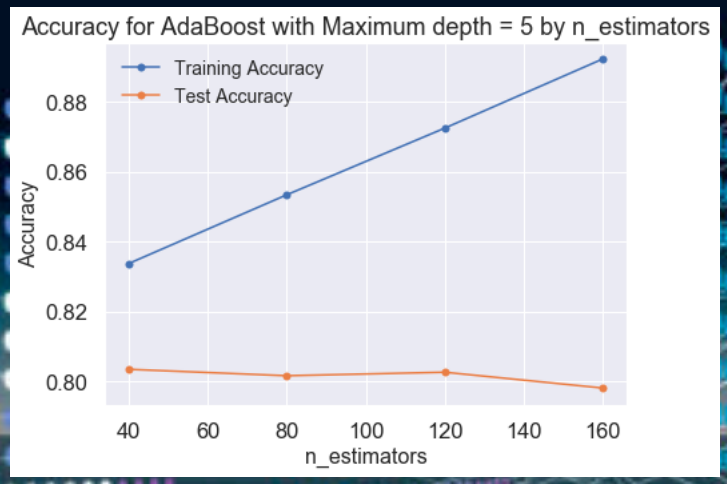


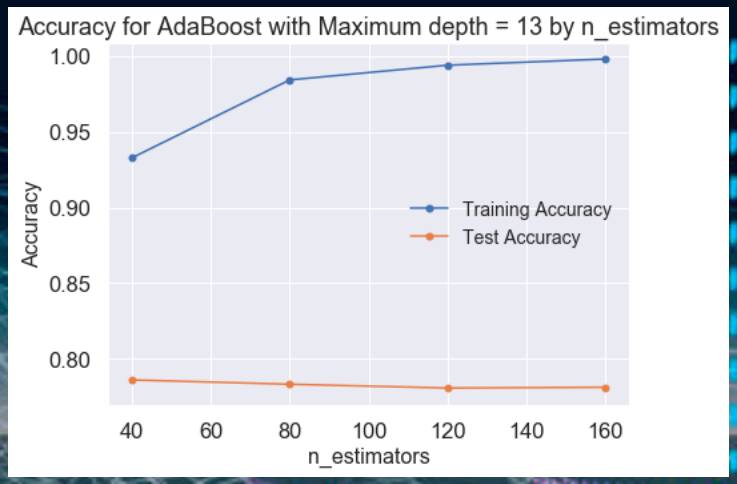
We chose max depth of 15 and Number of trees = 100

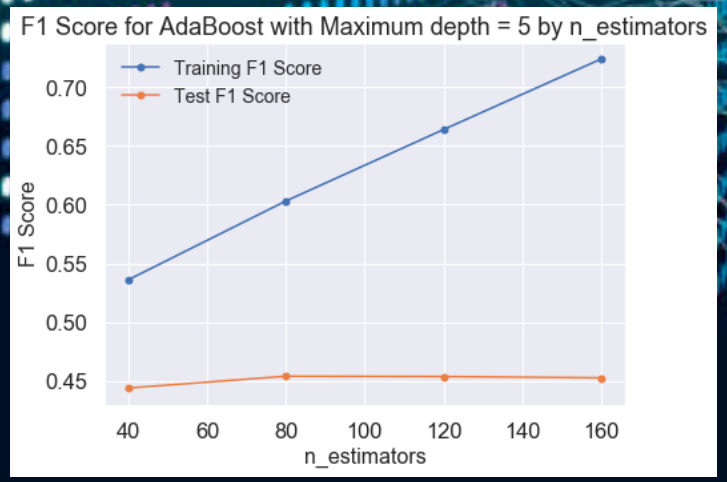
What is the random forest feature importance?

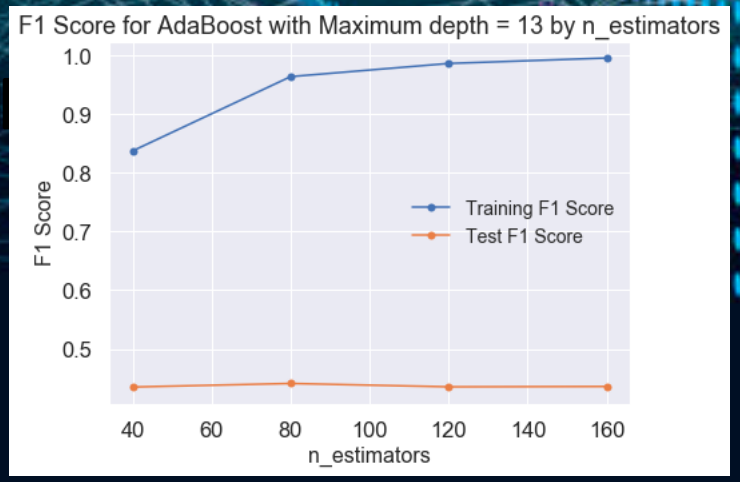


Ada Boost



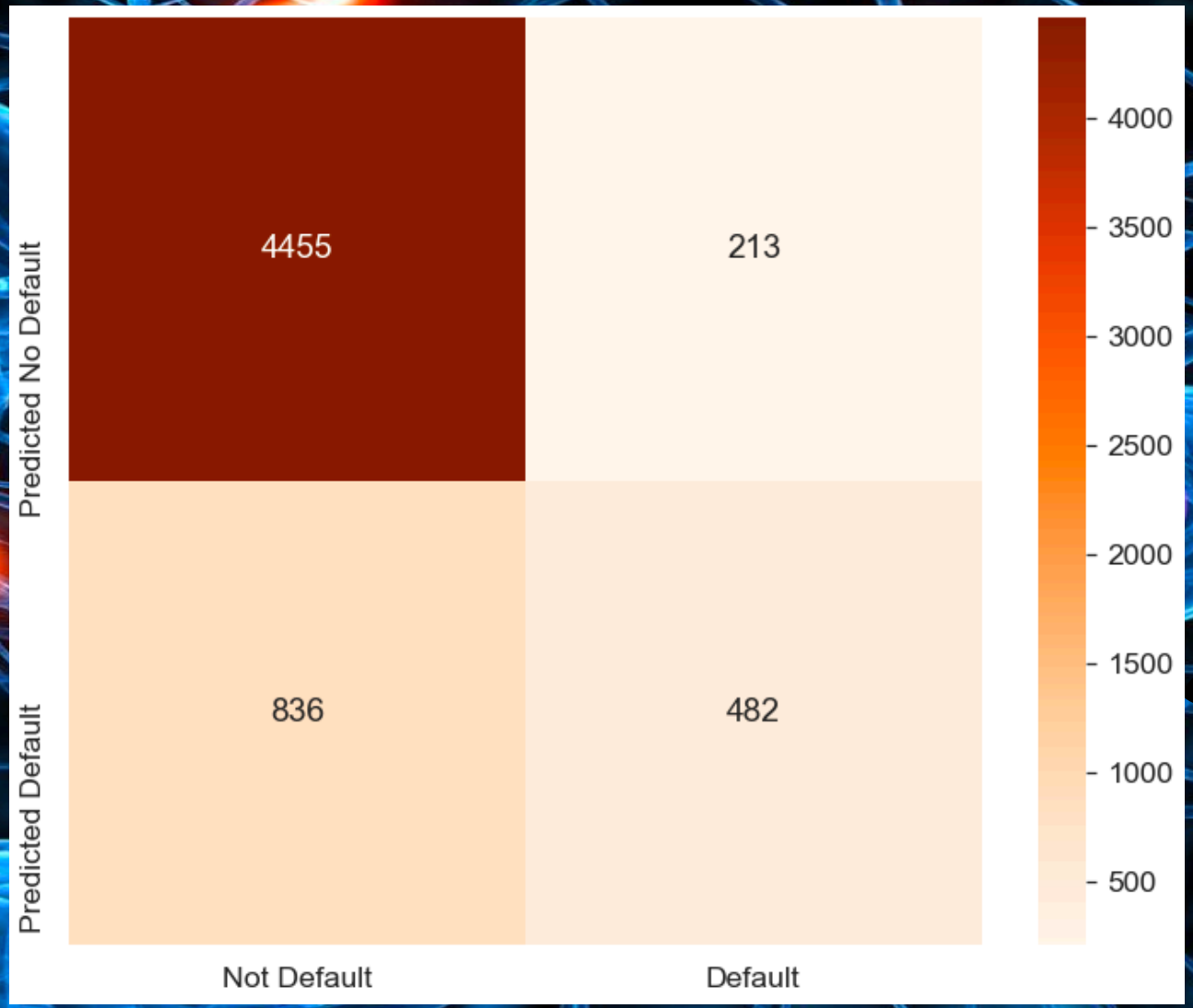






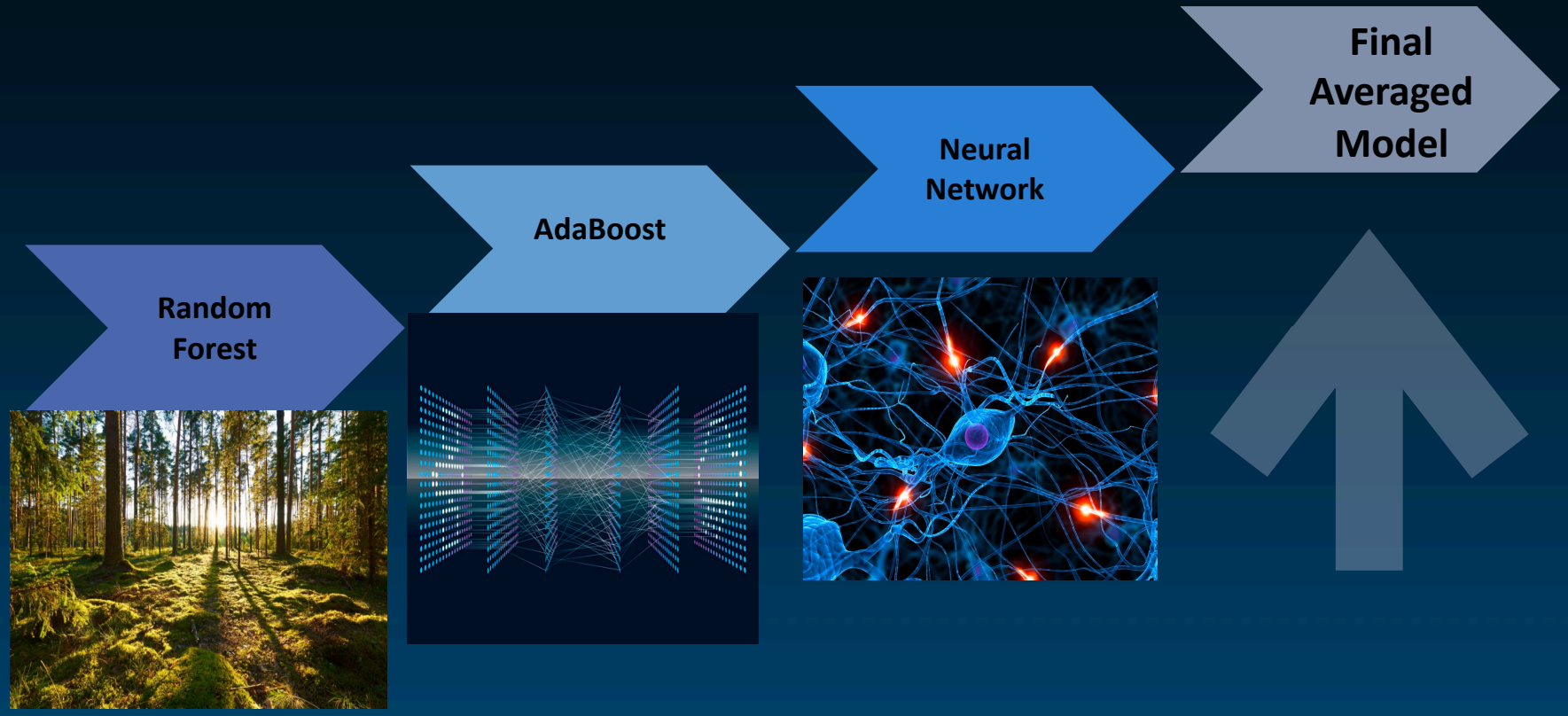
We chose max depth of 5 and Number of trees = 40

Neural network

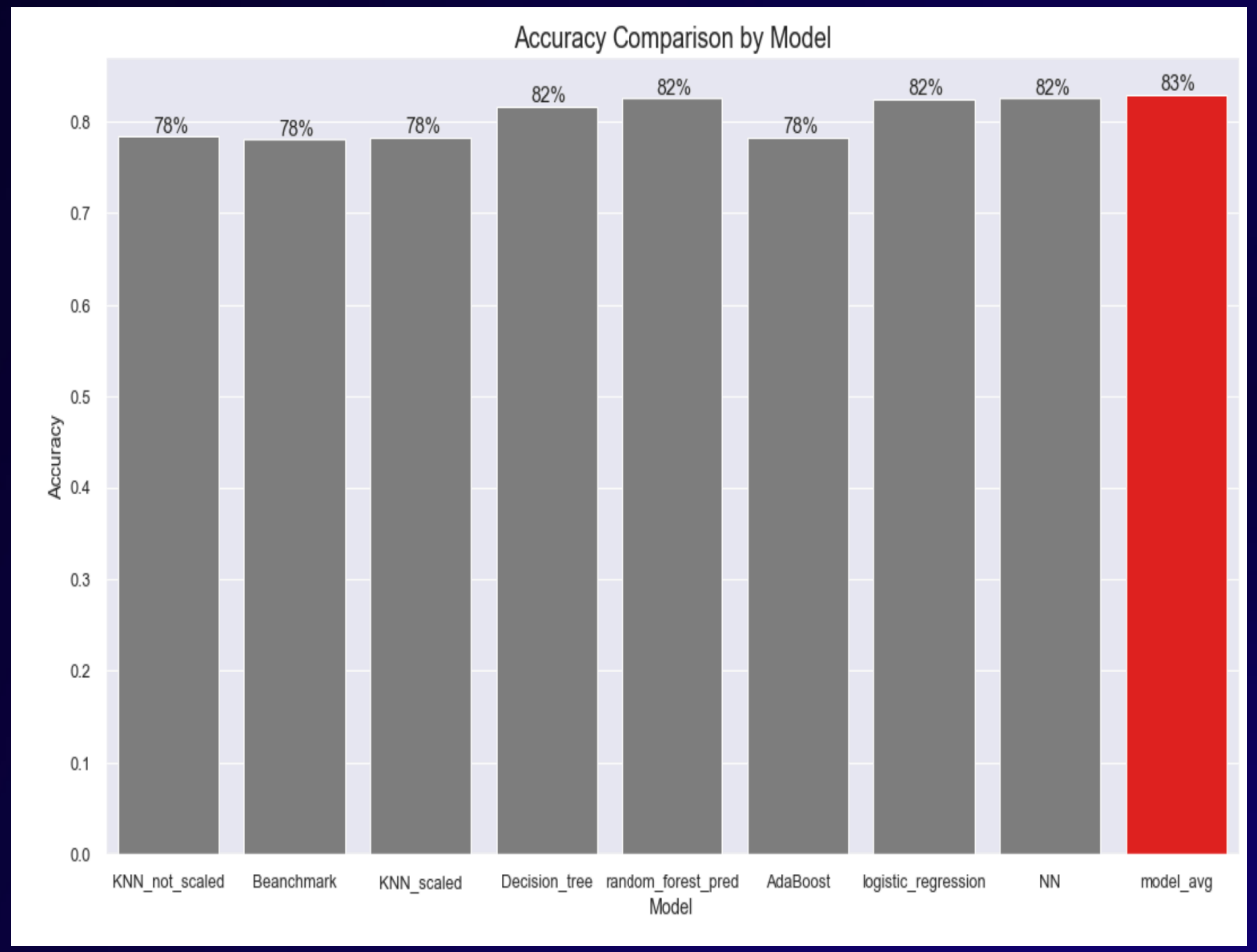


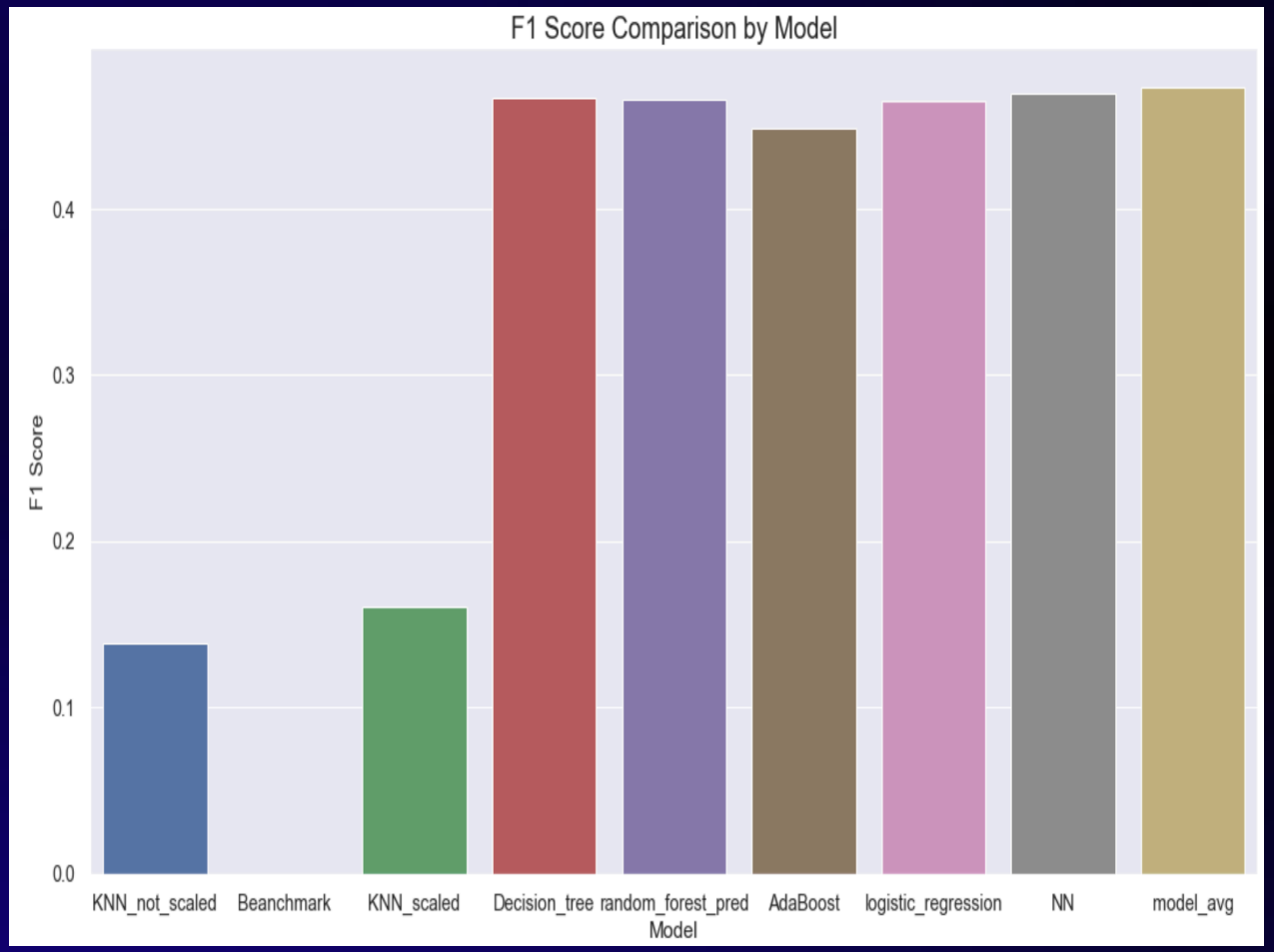
Accuracy of 82%
F1 Score of 45%

Averaging models

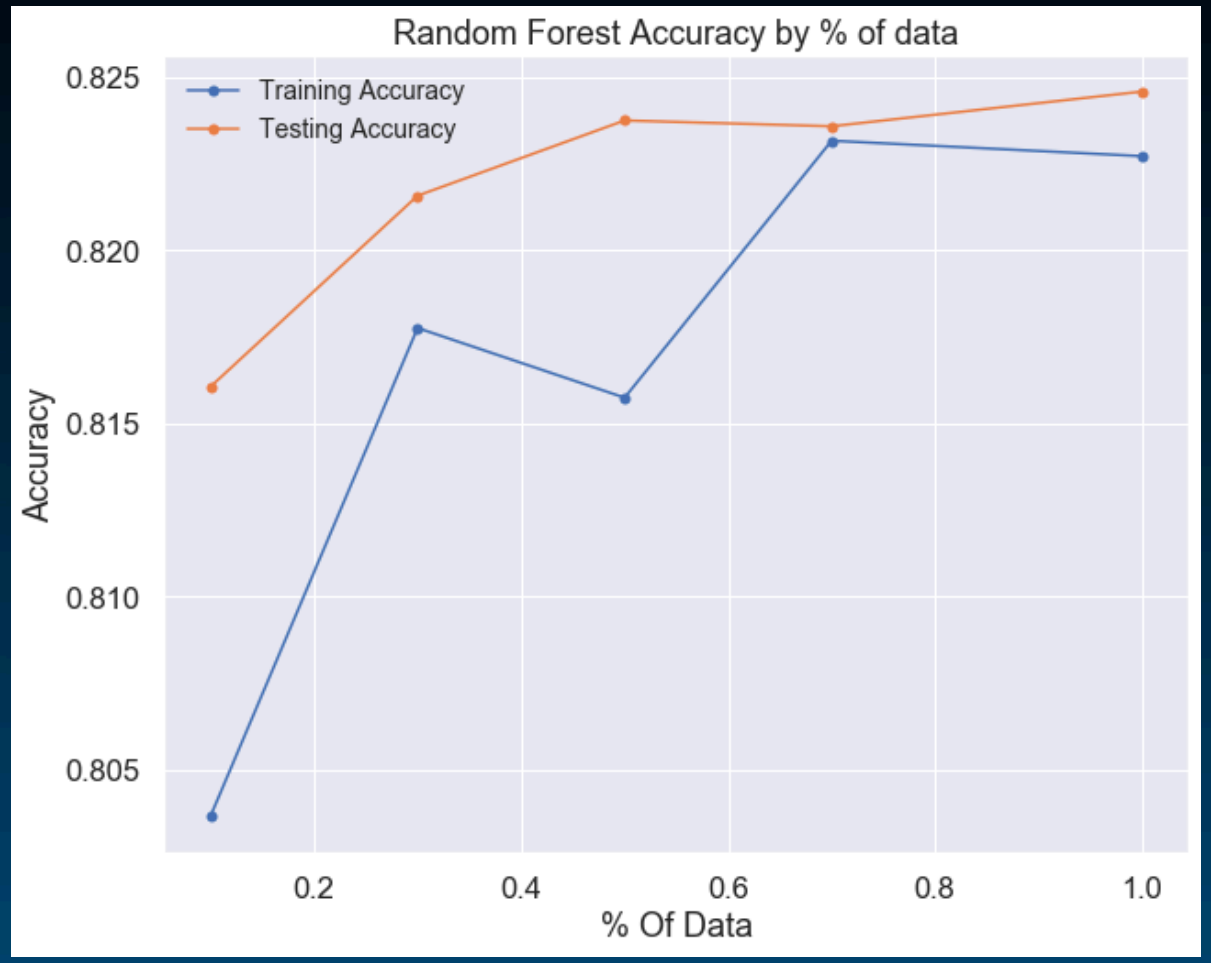


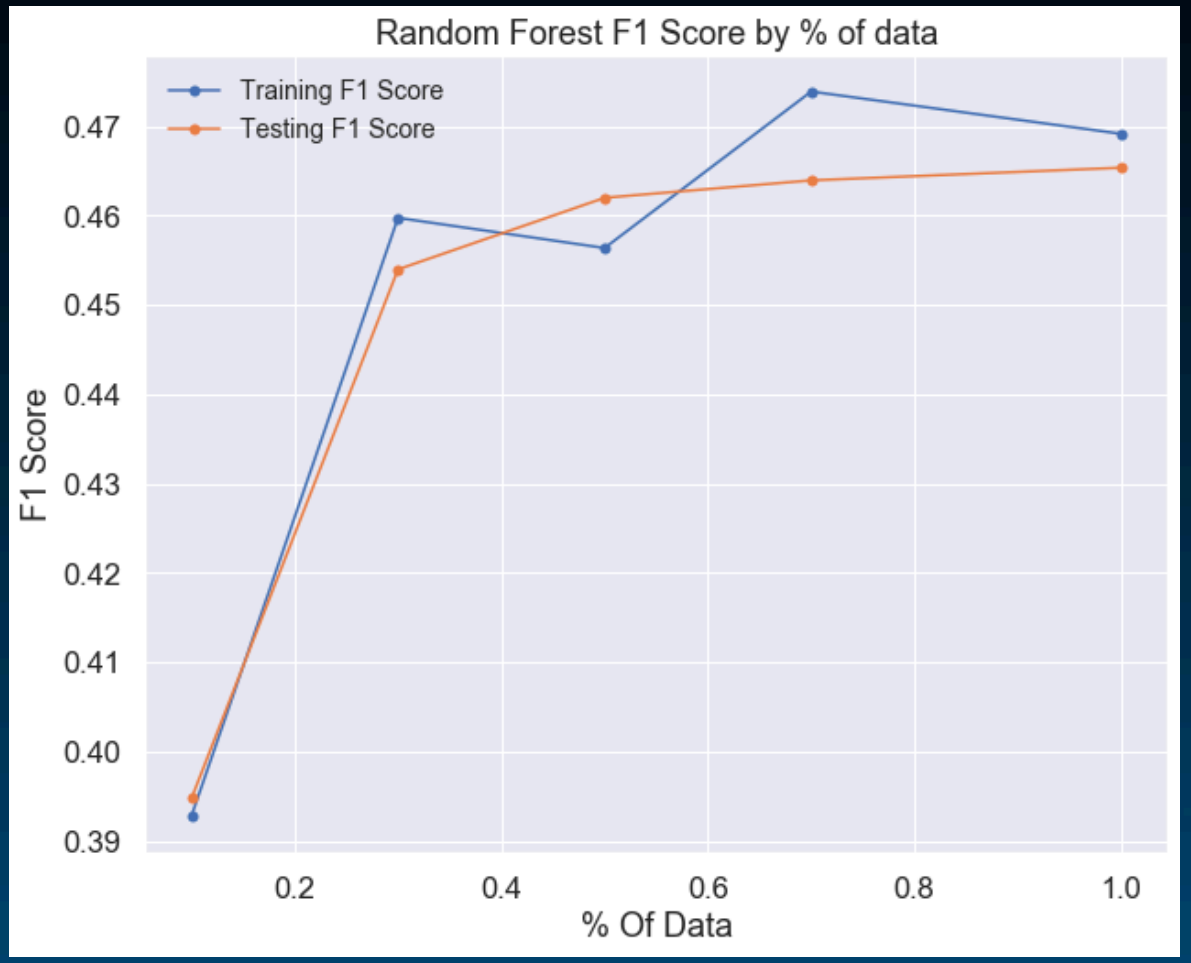
Averaging models





Performance vs. amount of data





It seems that there is no need for more data


THANK YOU