Classification Project

Catch the Extrovert

Nir Tal, Vitaliy Yashar

NAYA College
Bootcamp 2021

Materials

All the material available on **GitHub**

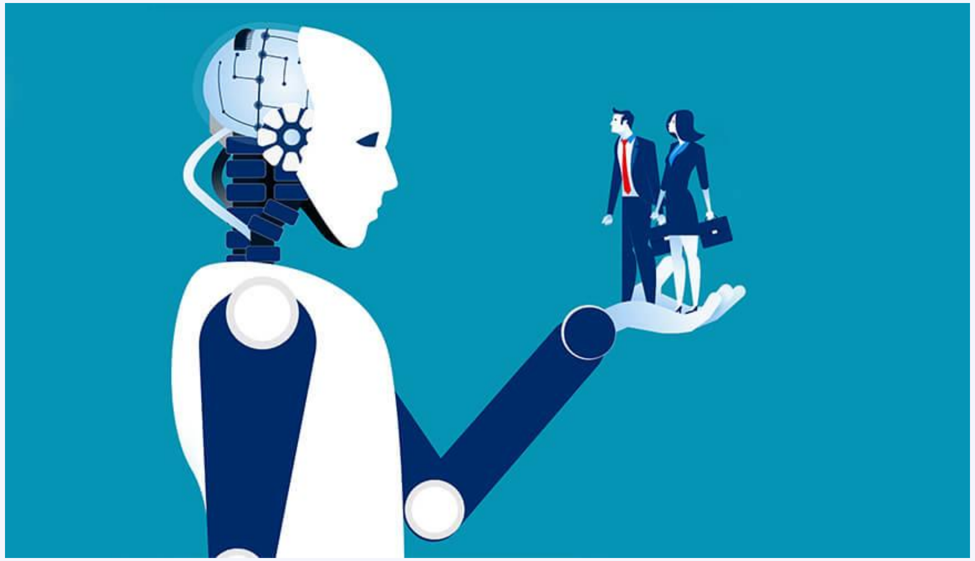
- The following presentation
- Jupyter notebooks
- Introvert/Extrovert Dataset
- “Pickled” Model

https://github.com/YasharDS

Introduction

Problem description

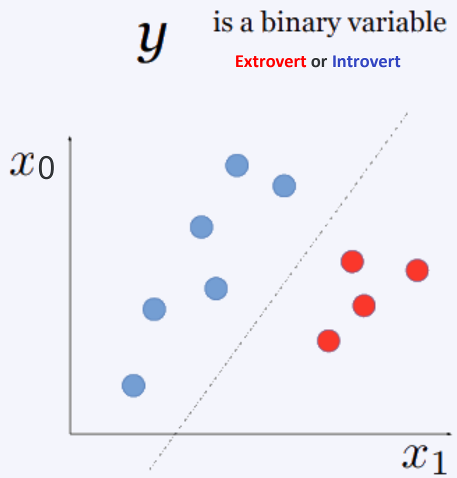
- **Mindspace HR department** is interested to hire Community Relation Manager and needs to point out the **extroverts(+)** from candidates list
- The firm is making an effort to select and continue the hiring process with **extroverts only**, as it is a very costly and time consuming hiring process



Proposed solution

- Given data set with ranking questionnaire and categories, will be introduced into classification algorithm in order to precisely classify the personality group

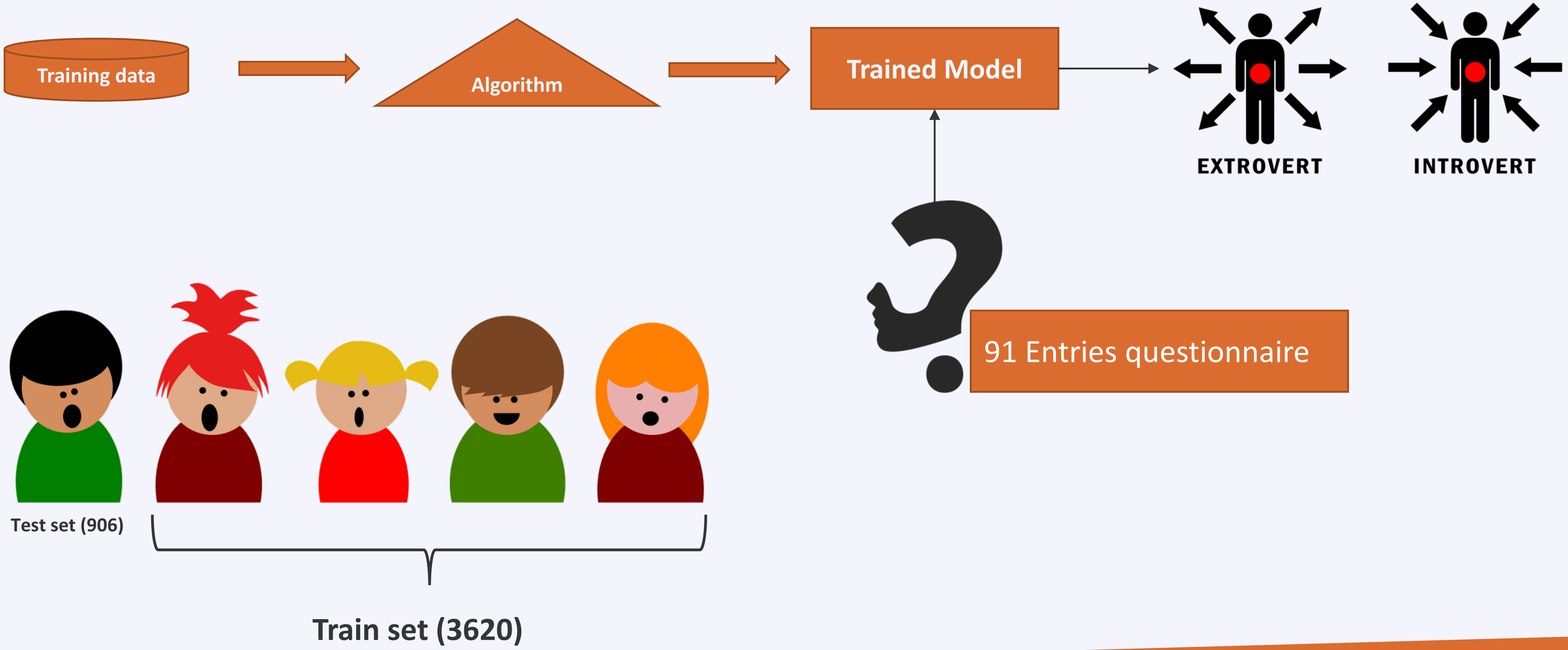
Machine learning tasks



Supervised learning

- Binary Classification: predict categorical **1(+)** for Extrovert **0(-)** for introvert
- Imbalanced data

Road Map



Training data

Algorithm

Trained Model

EXTROVERT

INTROVERT

91 Entries questionnaire

Test set (906)

Train set (3620)

Methodology

Supervised learning Binary Classification problem

Select classification model:

- Naive Base
- Support Vector Machines
- Random Forest
- Logistic Regression
- XGBoost

Train model, i.e., determine parameters

- Data: input + output
  - Training data → determine model parameters
  - Model improvements techniques

Model selection

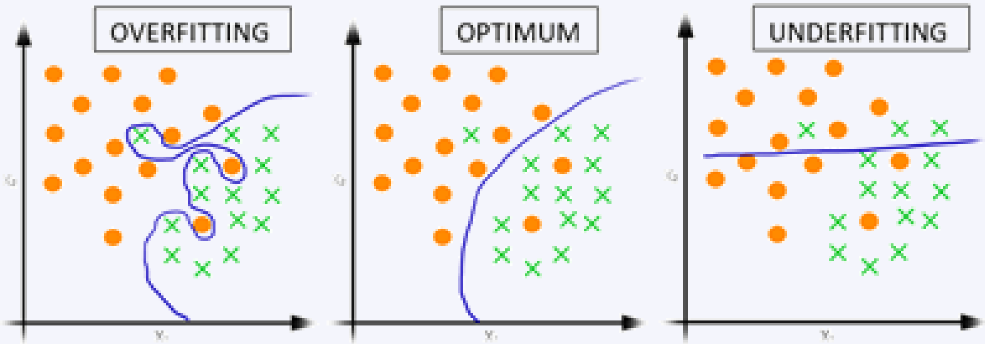
- Select the best model scored the highest AUC
- Fine-tune and Evaluate the best model by Precision and Recall

Test model

- Data: input + output
  - Testing data → final scoring of the model

Production

- Data: input → predict output



Classical Logistic Regression problem

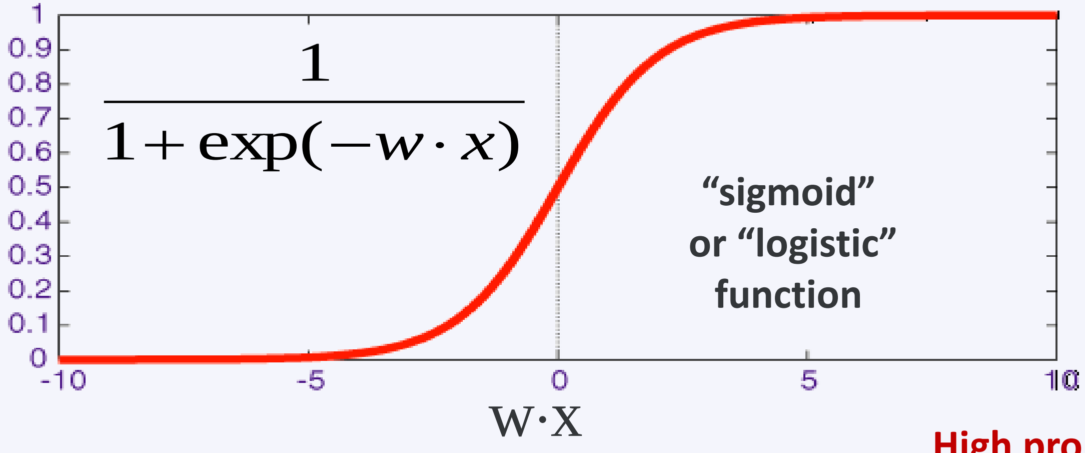
Logistic regression is derived from the following assumption:

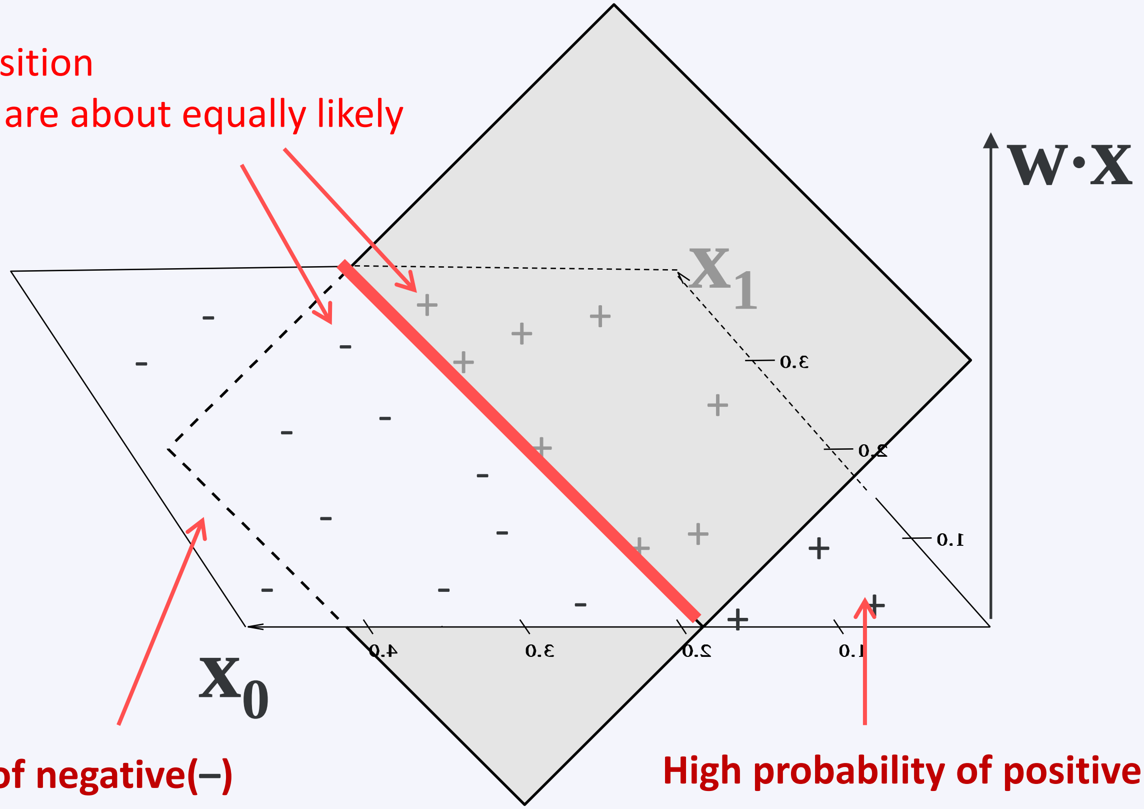
1. Suppose a true linear boundary exists, but is not a separator.
It caused the + and – labels to be assigned probabilistically

2. (W determines the boundary line and the gradualness of the transition)

3. Near boundary, a transition region where +’s and –’s are about equally likely

Probability that x is labeled (+)





High probability of negative(–)

High probability of positive(+)

Performance Metrics

Accuracy will not be enough to assess performance

$\textit{accuracy} = \frac{TP+TN}{P+N}$

Percentage of correctly classified instances.

$\textit{recall} = \frac{TP}{TP+FN}$

Ability of a model to find all the (+)positive cases within a dataset

$\textit{precision} = \frac{TP}{TP+FP}$

Fraction of relevant (+) instances among the selected ones

$F\ 0.5 = (1+\beta^2)\ \frac{2*presicion*recall}{(\beta^2*precision)+recall}$

Example of the Fbeta-measure with a beta value of 0.5
It has the effect of **raising the importance of precision** **and** **lowering the importance of recall**

False Predictions Prioritization

0-Introvert

Loose potential candidate

1-Extrovert

Continue with wrong candidate

Spending funds on hiring process

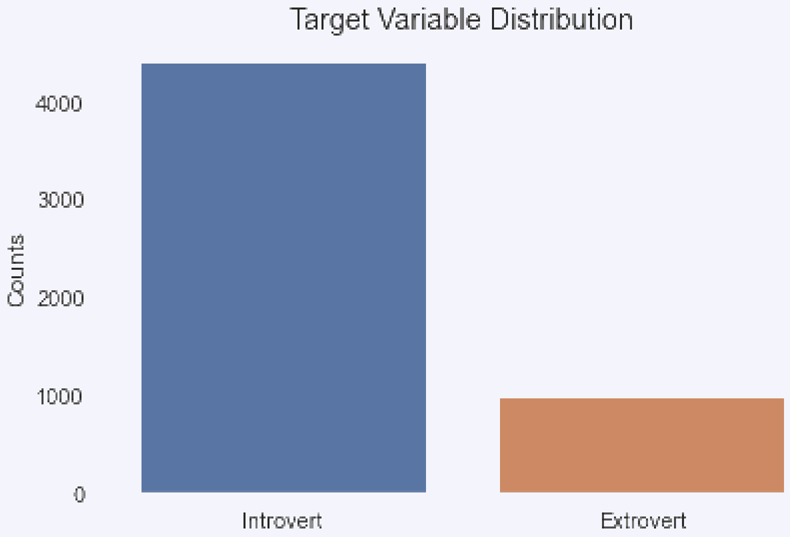
Loosing relevant potential candidates

Going back to additional hiring process

		Predicted Class	
		“negative” C=0	“positive” C=1
Actual	Y=0	TN	FP ⬇
	Y=1	FN ★	TP

Increase the Importance of Precision

Data Cleaning and Preparation



0-Introvert (-) **1-Extrovert (+)**

Dataset for Training and Testing

- **Each person has number of answers and additional info regarding the exam**
- **The responses for each question ranked between 1-5 (Disagree-Agree)**
- **Each person identifies himself as either introvert or extrovert**
- 4526 entries after processing [3713(-), 813(+)]

Several Data cleaning steps implemented:

- All of the variables were cleaned, except of the 91 questions
- This step reduced the features from 282 to 91 + Target Var.

By the end of the process, it's clear that we stand in front of imbalanced data

- Introverts(-) class is 4.6 time more frequent than the Extrovert(+)

Data Wrangling

- <u>Data ingestion</u>

  CSV Data set (7188 entries and 282 features)

  Male/Female is not relevant as seen from the distribution

  Same as “Is English your mother tongue”

  We assume that the age would not provide us additional insight

  If the personality group is not introvert or extrovert → *remove*

  1. After this step : 4404 entries(Introvert), 990 entries(Extrovert)

- <u>Data cleaning</u>

  Outliers/invalid values? → filter

  1. *The tests that took more than **900sec** to fill up → removed*
  2. *The last page time>**50 sec**→ removed*

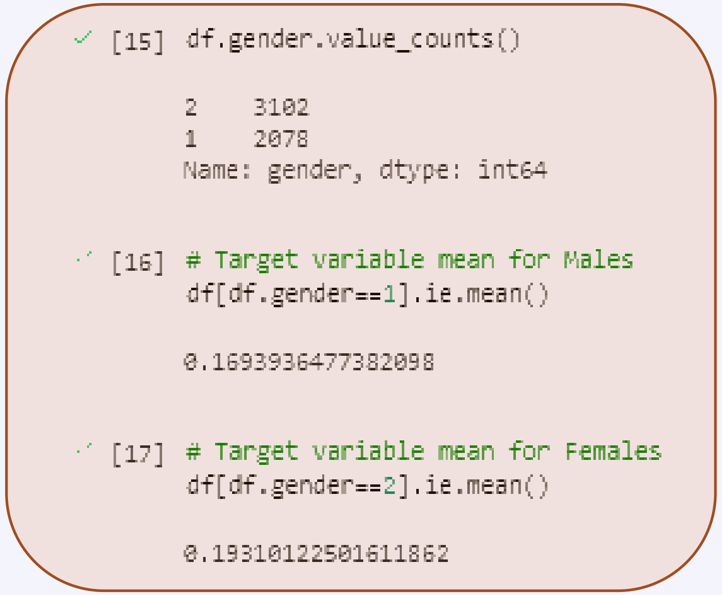
  3. The Extrovert will indicate positive “1” in this project

  Missing values? → impute or remove

  1. No missing Values observed

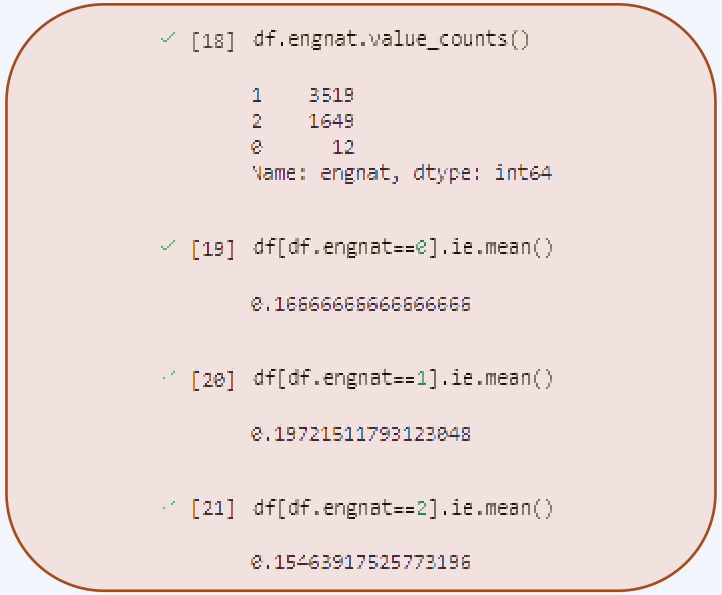
- <u>Data transformation</u>

  Scaling/normalization → Not necessary. Data structured well



```
✓ [15] df.gender.value_counts()

       2    3102
       1    2078
       Name: gender, dtype: int64

✓ [16] # Target variable mean for Males
       df[df.gender==1].ie.mean()

       0.1693936477382098

✓ [17] # Target variable mean for Females
       df[df.gender==2].ie.mean()

       0.19310122501611862
```

Male/Female Distribution



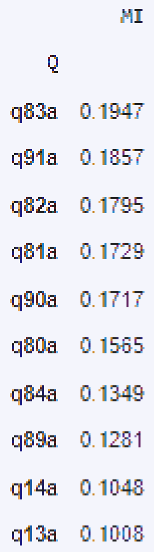
```
✓ [18] df.engnat.value_counts()

       1    3519
       2    1649
       0      12
       Name: engnat, dtype: int64

✓ [19] df[df.engnat==0].ie.mean()

       0.16666666666666666

✓ [20] df[df.engnat==1].ie.mean()

       0.19721511793123048

✓ [21] df[df.engnat==2].ie.mean()

       0.15463917525773196
```

English mother tongue Distribution

Features Relevancy

At the first place we tried to identify (**Select_Kbest** & **mutual_info_classif)**

the most relevant question, by choosing only 9-12 questions which scored the highest scores



	MI
Q	
q83a	0.1947
q91a	0.1857
q82a	0.1795
q81a	0.1729
q90a	0.1717
q80a	0.1565
q84a	0.1349
q89a	0.1281
q14a	0.1048
q13a	0.1008

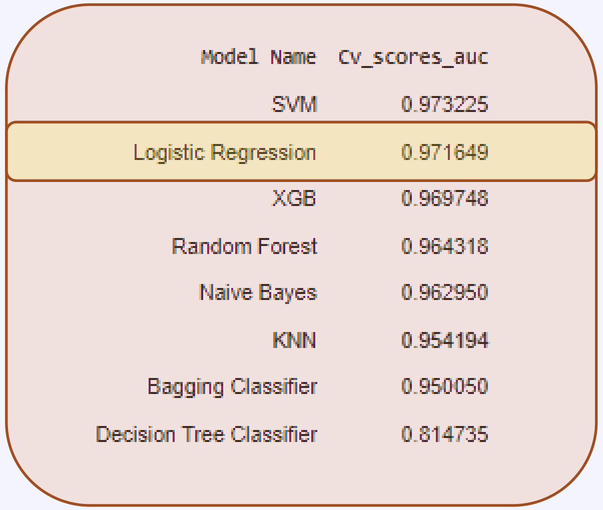
After training the models and evaluating them with ROC_AUC, we retrained them on **the full data set** (91 questions)

The models scored better:

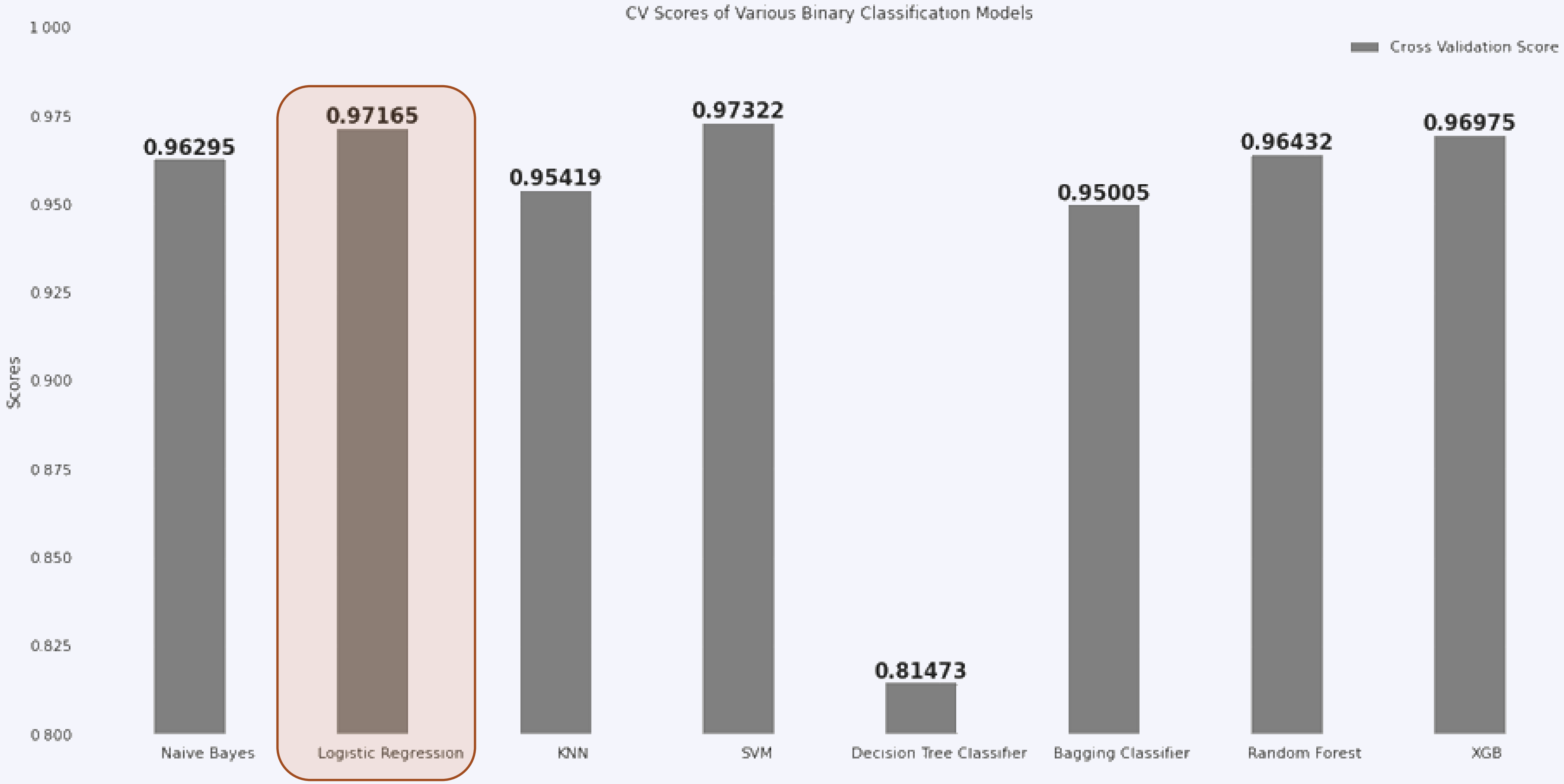
ROC_AUC of LogisticRegression increased from 0.964 to 0.973

Cross Validated Models Evaluation by ROC_AUC

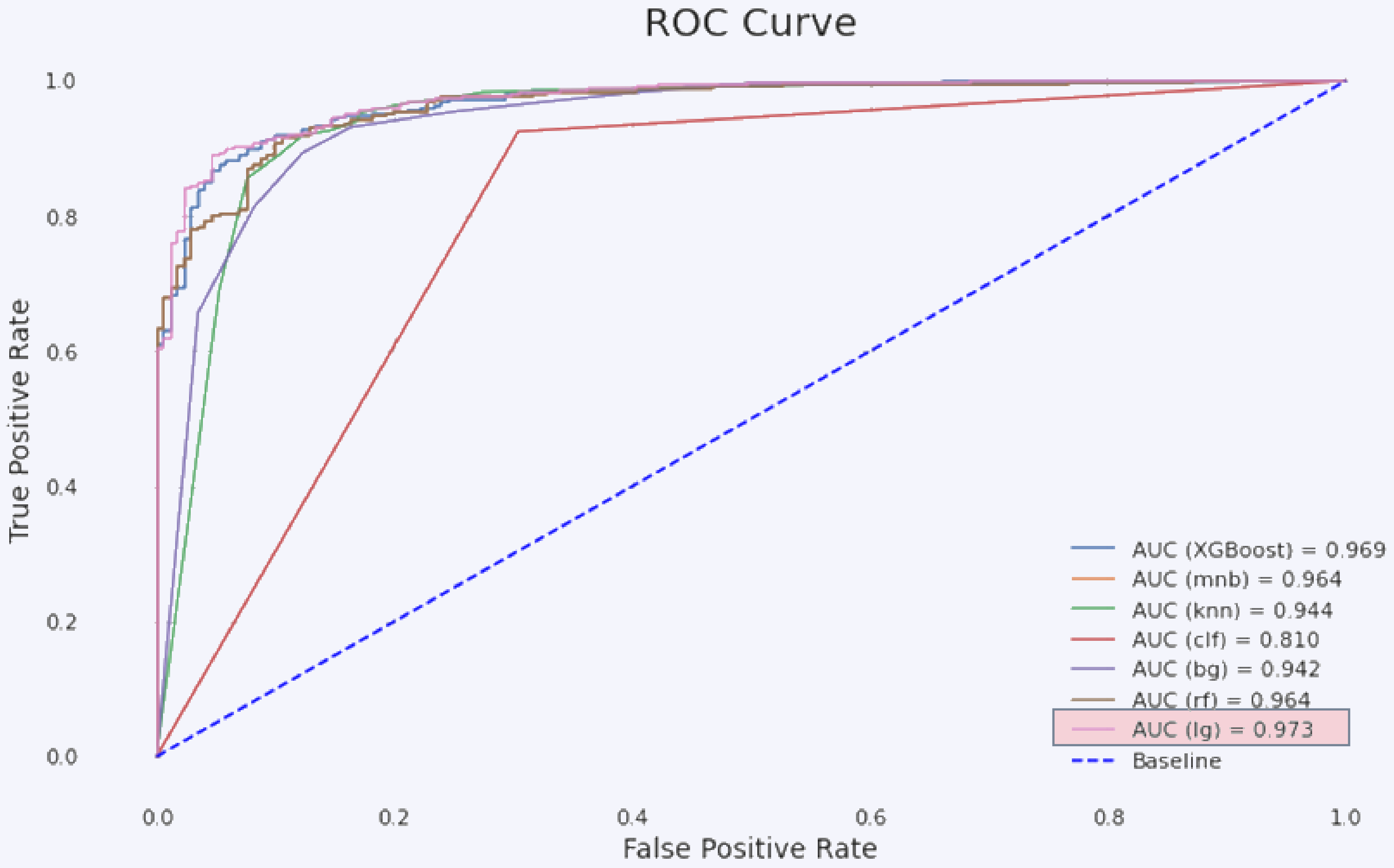
- From initial evaluation, we conclude that **LR model scored highest** and was valid
- Cross validating was capable of improving the SVM score
- There is a room for improvement of decision tree, but we would not further research it, in terms of this problem
- The logistic regression scored pretty the same as SVM, but we **decided to continue with Logistic Regression** as it classic for binary classification and it will be easily introduced to the higher management



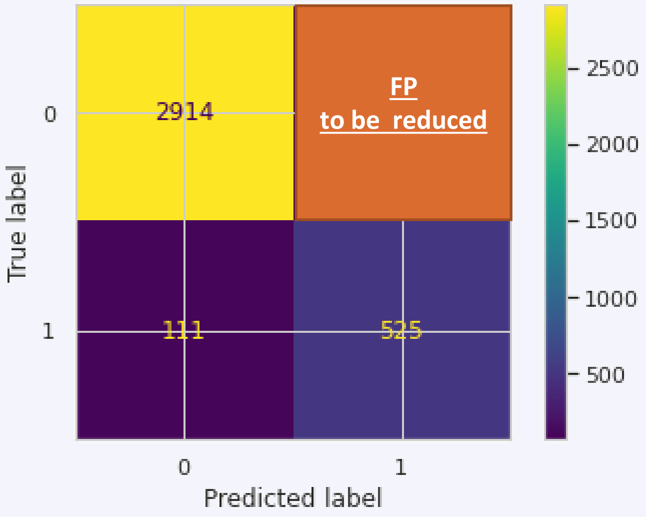
Model Name	Cv_scores_auc
SVM	0.973225
Logistic Regression	0.971649
XGB	0.969748
Random Forest	0.964318
Naive Bayes	0.962950
KNN	0.954194
Bagging Classifier	0.950050
Decision Tree Classifier	0.814735



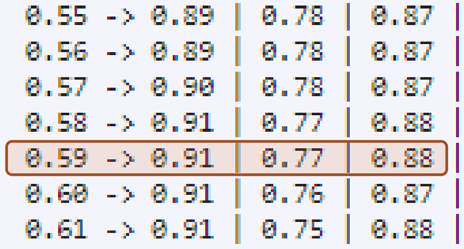
Cross Validated Models Evaluation ROC Curves



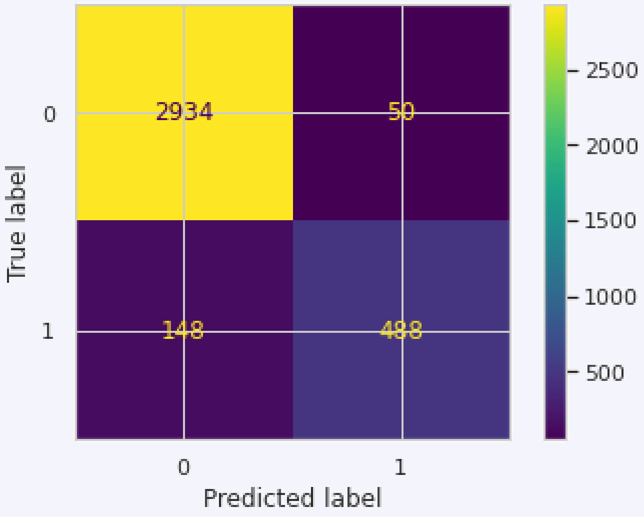
Confusion Matrix and Logistic Regression Improvement



Tresh Prec Recall F0.5



```
0.55 -> 0.89 | 0.78 | 0.87 |
0.56 -> 0.89 | 0.78 | 0.87 |
0.57 -> 0.90 | 0.78 | 0.87 |
0.58 -> 0.91 | 0.77 | 0.88 |
0.59 -> 0.91 | 0.77 | 0.88 |
0.60 -> 0.91 | 0.76 | 0.87 |
0.61 -> 0.91 | 0.75 | 0.88 |
```



First Model

Threshold	0.50
Precision	0.88
Recall	0.83
F0.5	**0.87**

Regularization

C	0.01
F0.5	**0.85±0.02**

Threshold selection

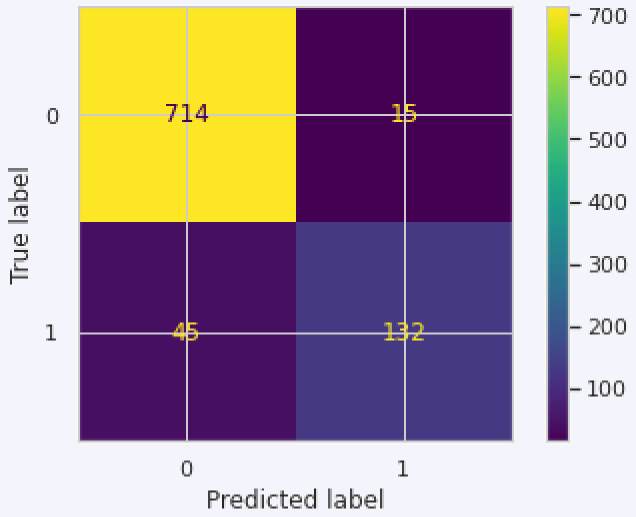
Threshold	0.59
Precision	0.91
Recall	0.77
F0.5	**0.88**

Summary and Model Test

Model Selection: LogisticRegression

(C=0.01, Threshold=0.59)

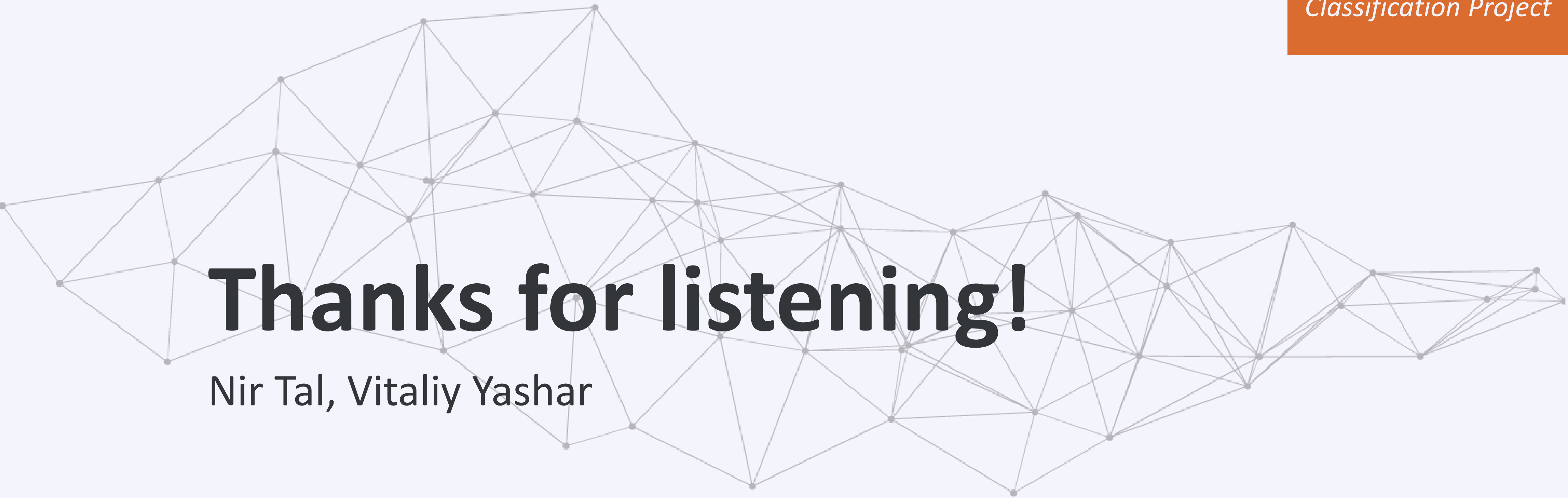
Metrics	**Yields**
Precision	0.90 (-0.01)
Recall	0.75 (-0.02)
F0.5	**0.86** (-0.02)



Test data Model Performance

LogisticRegression: LG.pkl

- Fine-tuning the model yielded a **LOWER FP rate**, which was the main goal
- **Threshold optimization** is one of the main actions to perform on imbalanced data
- In conclusion, the LR model scores the best when dealing with Binary Classification problems, when the features are from the same scaling system



Thanks for listening!

Nir Tal, Vitaliy Yashar

NAYA College
Bootcamp 2021