Classification Project

Catch the Extrovert

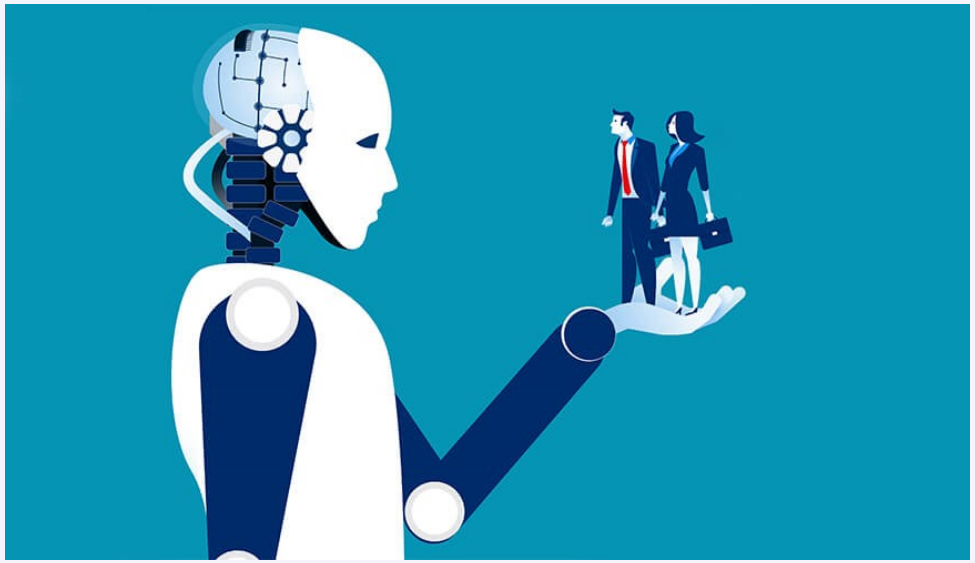
Nir Tal & Vitaliy Yashar

NAYA College
DS Bootcamp

Introduction

Problem statement

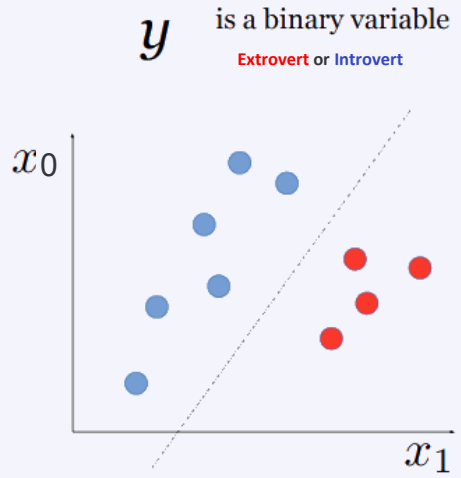
- Mindspace is interested in hiring a **Community Manager** and needs to pick the **extroverts** from the candidate list
- The company's goal is to start the hiring process with **extroverts** only, because the process is very costly and time consuming



Proposed solution

- Based on the MIES online personality test, we will develop a model for the classification of candidates into two personality groups: introverts and extroverts

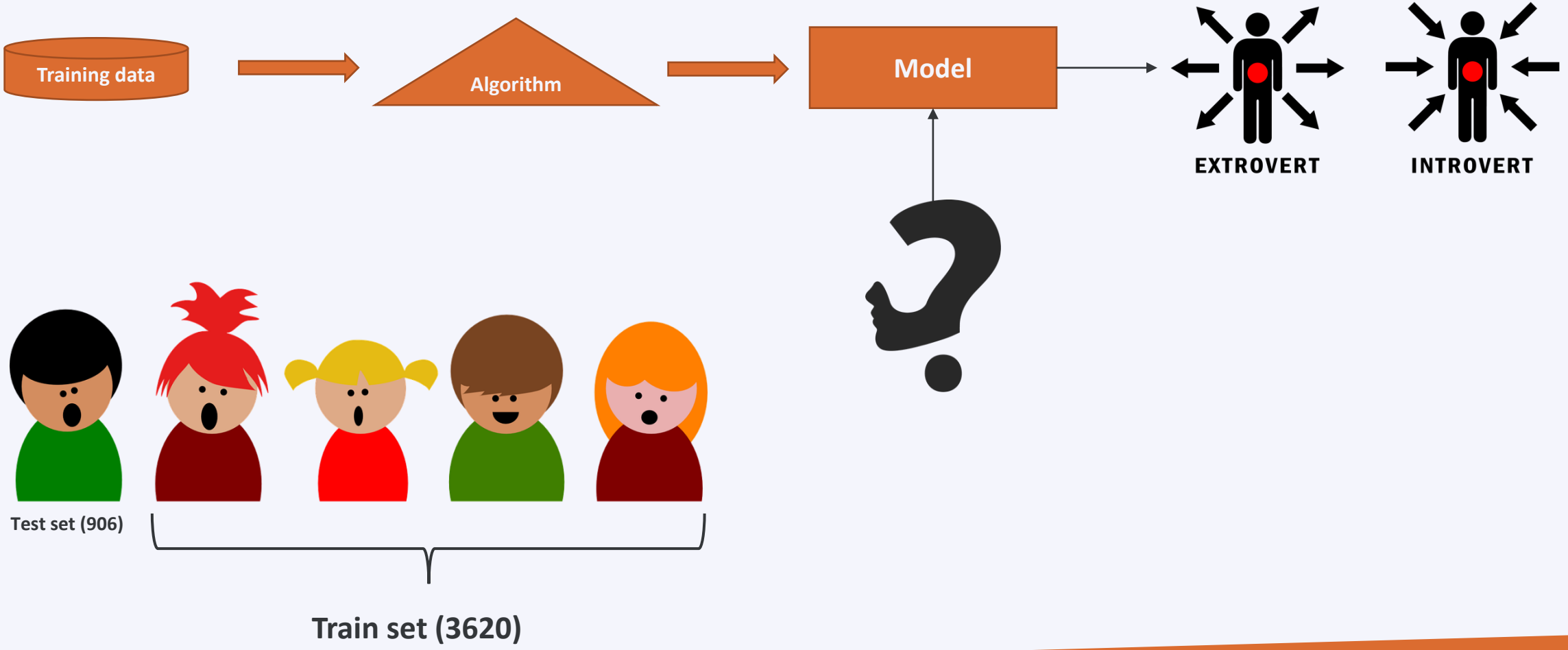
Machine Learning Tasks



Supervised Learning

- Binary Classification: categorical **1** for extroverts or **0** for introverts
- Caveat: imbalanced data

Road Map


Training data
Algorithm
Model
EXTROVERT
INTROVERT
Test set (906)
Train set (3620)

Methodology

Supervised Learning Binary Classification Problem

Select classification model:

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

Train model & determine parameters

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

Test model

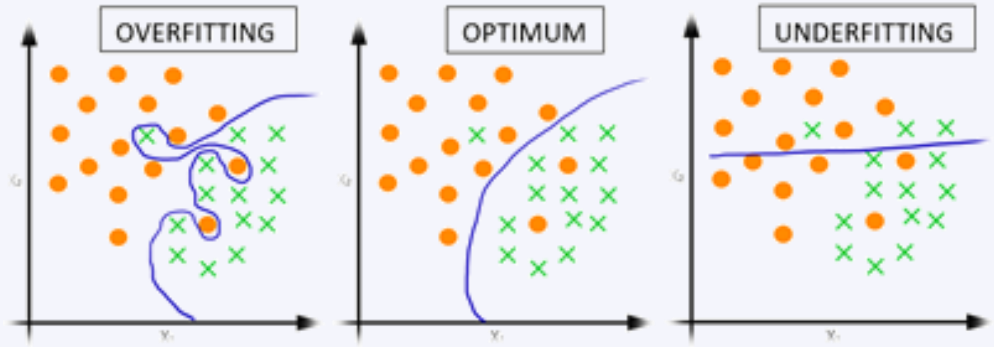
- Data: input + output
  - Testing data → final scoring of the model

Model selection

- Select best model according to highest AUC score
- Fine-tune and evaluate model according to Precision and Recall scores

Prediction

- Data: input → predict output



Logistic Regression Problem

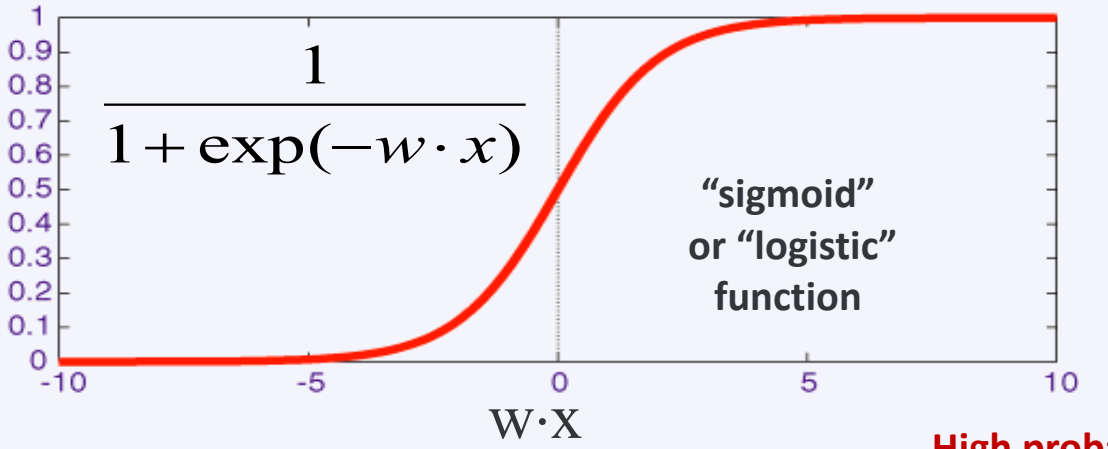
- Logistic regression is derived from the following assumption:

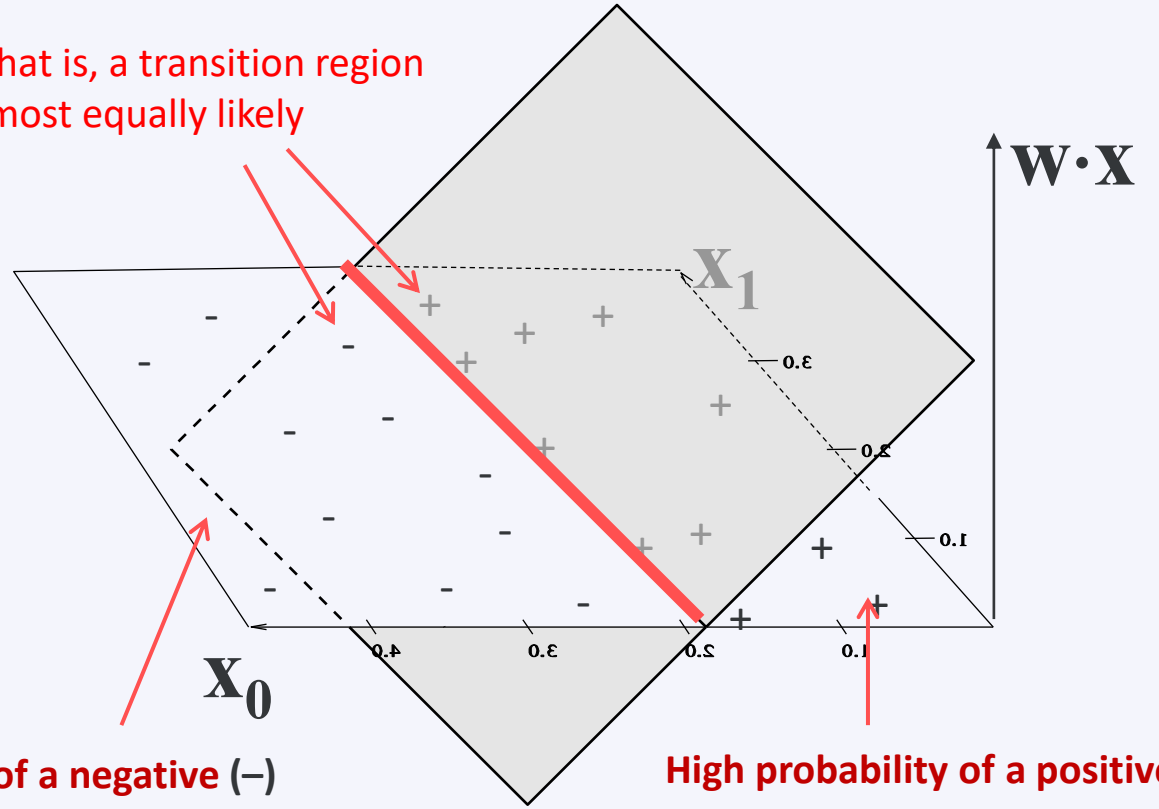
- Suppose a <u>true</u> linear boundary exists, but it's not a separator.
  It causes the + and – labels to be assigned probabilistically

W determines the boundary line and the gradualness of the transition

A very close boundary, that is, a transition region where (+) and (–) are almost equally likely

The probability that x is labeled (+)





High probability of a negative (–)

High probability of a positive (+)

Performance Metrics

Accuracy will not be enough to assess performance

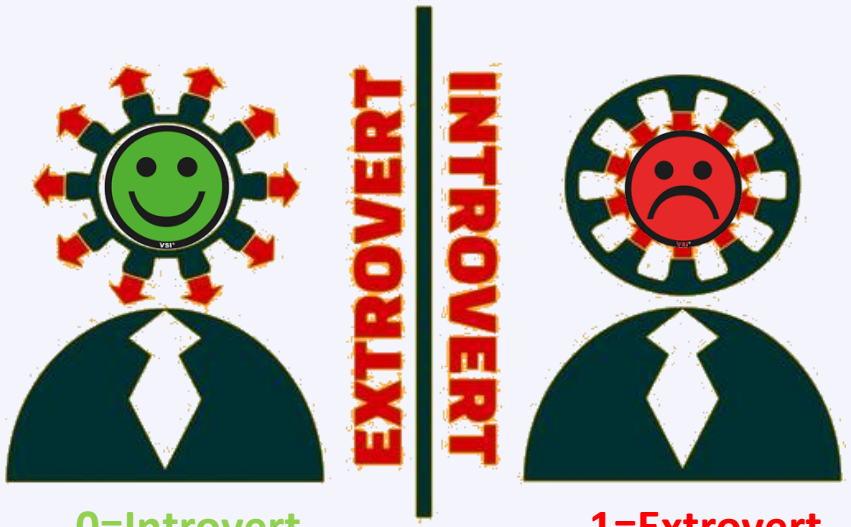
$Accuracy = \frac{TP+TN}{P+N}$ Percentage of correctly classified instances

$\boldsymbol{Recall} = \frac{TP}{TP+FN}$ Ability of a model to find all the positive cases within a dataset

$\boldsymbol{Precision} = \frac{TP}{TP+FP}$ Fraction of relevant (+) instances among the selected ones

$\boldsymbol{F0.5} = (1+\beta^2)\frac{\boldsymbol{presicion}*\boldsymbol{recall}}{(\beta^2*\boldsymbol{precision})+\boldsymbol{recall}}$ 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 Trade-off



0=Introvert

Loose a potential candidate

1=Extrovert

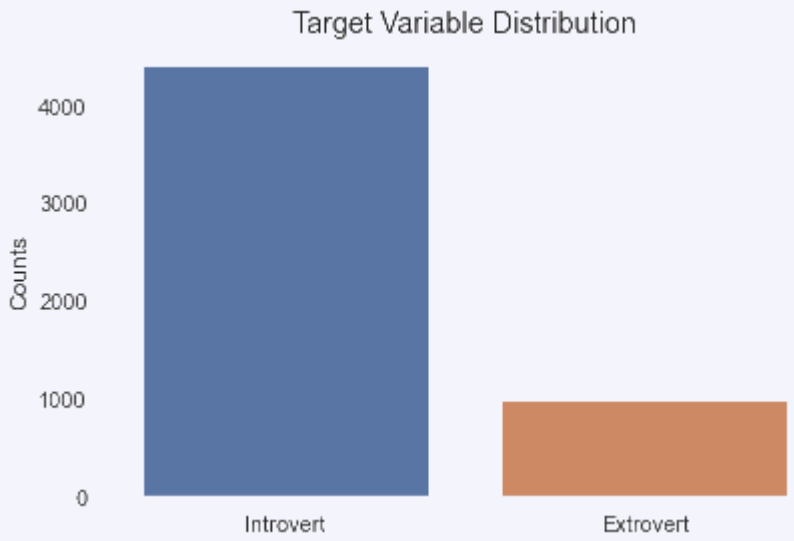
Continue with the wrong candidate

Waste time & money on hiring process

		Predicted Class	
		"negative" C=0	"positive" C=1
Actual	Y=0	TN	FP ↓
	Y=1	FN ★	TP

We should increase the importance of Precision!

Data Cleaning and Preparation



Introvert (0) **Extrovert (1)**

Dataset for training and testing

- Each person has several answers and additional info regarding the exam
- Responses for each question are ranked between 1=Disagree to 5=Agree
- Each person identifies himself as either Introvert (0) or Extrovert (1)
- 4526 entries after processing: 3713 (0) & 813 (1)

Several data cleaning steps were implemented:

- All variables were removed, except for the 91 questions
- This step reduced the features from 282 to 91

By the end of the process, it's clear that we face an imbalanced data

- The Introvert class is 4.5 times more frequent than the Extrovert class

Data Wrangling

Data ingestion

- CSV data set (7188 entries and 282 features)
- Gender is not relevant as seen from distribution
- English (as native language) is not relevant
- We assume that Age is not relevant
- If personality group is not introvert or extrovert → *remove*
  After this step: 4404 Introvert *entries & 990* Extrovert *entries*

Data cleaning

Outliers/invalid values? → filter

- *Exams that took more than **900 sec** to answer → remove*
- *Last page dwelling time > **50 sec** → remove*
- ***Extrovert** is the **positive label (1)** in this project*

Missing values? → impute or remove

- *No missing values observed*

```
[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
```

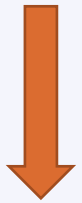
Male/Female distribution

```
[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 as native language distribution

Feature Selection

First we selected the 12 most important questions (i.e., features)
Using the **mutual_info_classif** method:

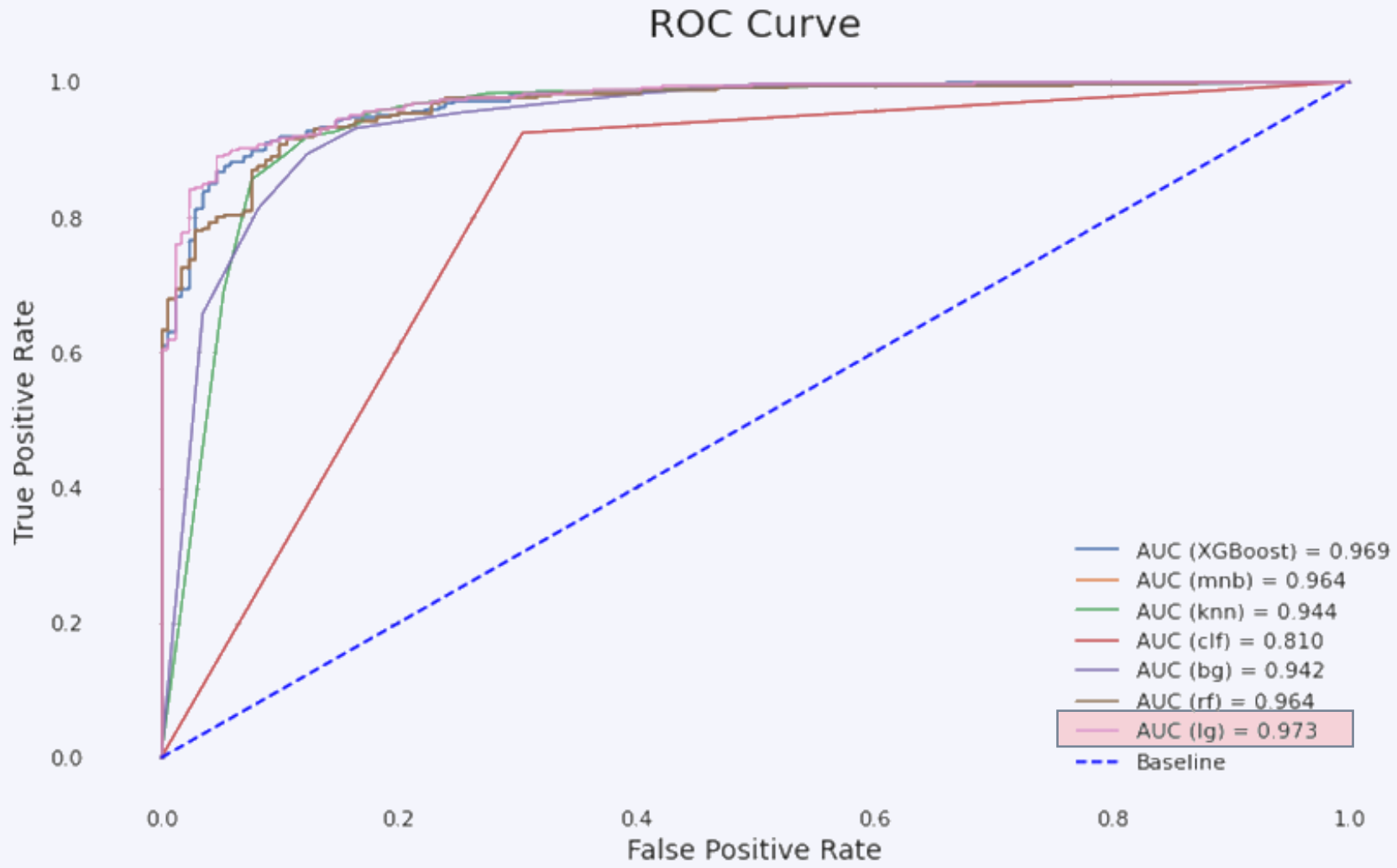


After training the models and evaluating them with the ROC_AUC metric;

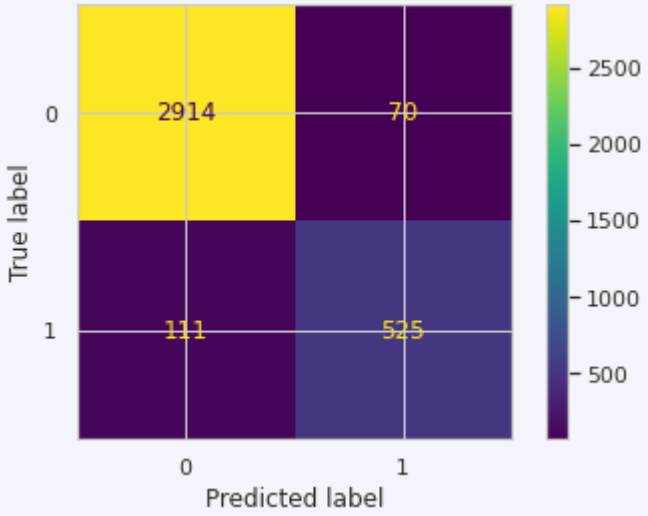
We retrained them on **the full data set** (91 questions):

ROC_AUC of Logistic Regression increased from 0.964 to 0.973

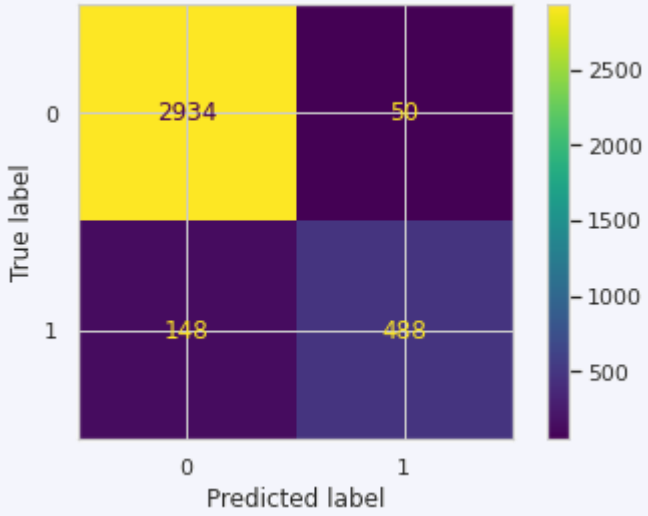
Classification Models: Evaluation with ROC_AUC



Logistic Regression Model & Confusion Matrix



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

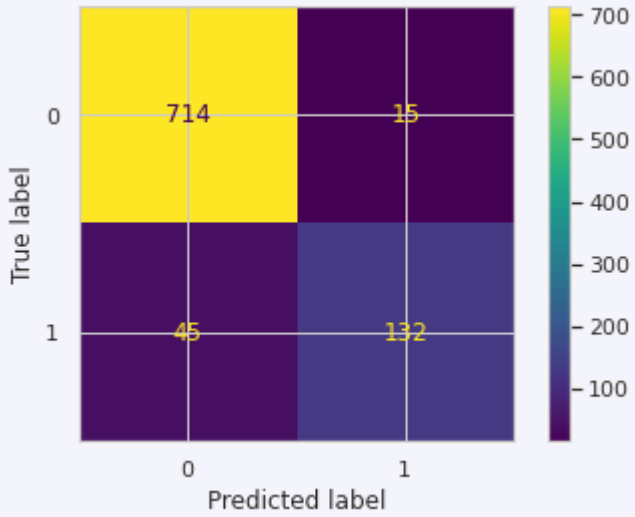
Threshold	**0.59**
Precision	0.91
Recall	0.77
F0.5	**0.88**

Logistic Regression Model: Testing Data

Model Selection: Logistic Regression

(C=0.01, Threshold=0.59)

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



Model performance on testing data

- Fine-tuning the model yielded a **lower FP rate**, which was the main objective
- **Threshold optimization** is one method to overcome imbalanced data
- In conclusion, the Linear Regression model scored the best for this binary classification problem

Classification Project

Thank You!

Nir Tal & Vitaliy Yashar

NAYA College
DS Bootcamp