Supervised Machine Learning

Modelling Customer Response on the Bank Marketing dataset

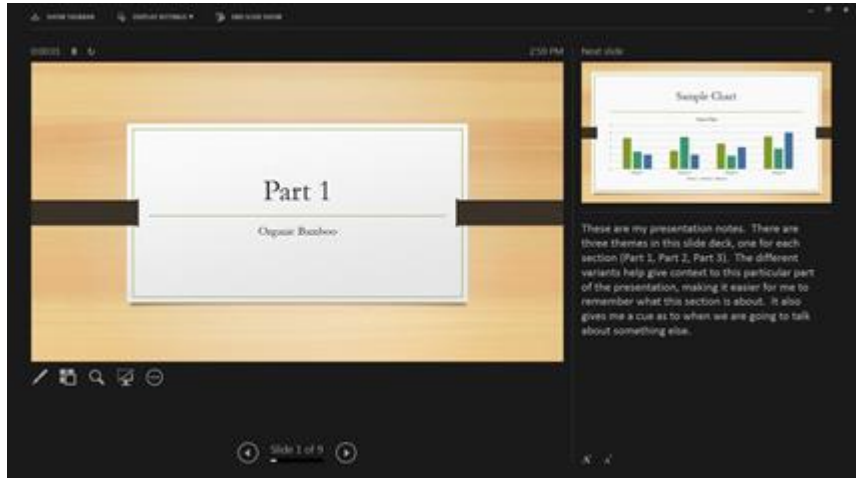
Presented by:
Sony Jufri

Business problem

Big marketing budget, poor outcome.

How to optimize marketing strategies and improve effectiveness by targeting the right customers.

Bank Marketing Dataset -> Machine Learning -> Predictive Modelling



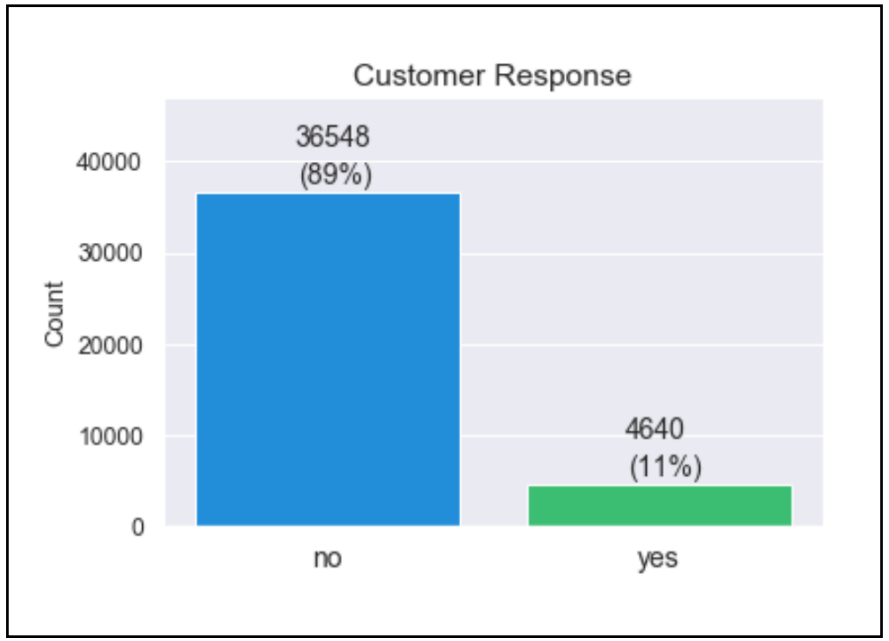
Bank Marketing Dataset

41,188 examples x 20 attributes

- Imbalanced dataset
- Missing values
- Outliers

Solution:

- **Cleaned**
- **Normalized**
- **Up-sampled**



python™

NumPy

pandas

matplotlib
Version 3.2.1

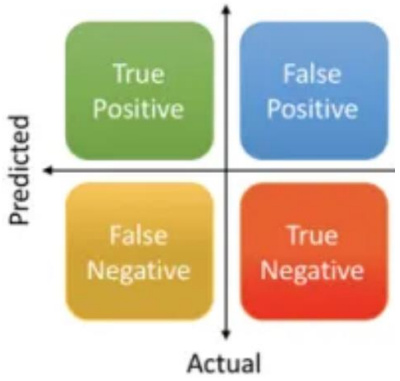
seaborn

Research Question & Hypotheses

Q: "Is the recommended Random Forest model better than the alternatives considered?"

H_0: The recommended model's f1-score is not statistically better than the alternatives

H_a: The recommended model's f1-score is statistically better than the alternatives



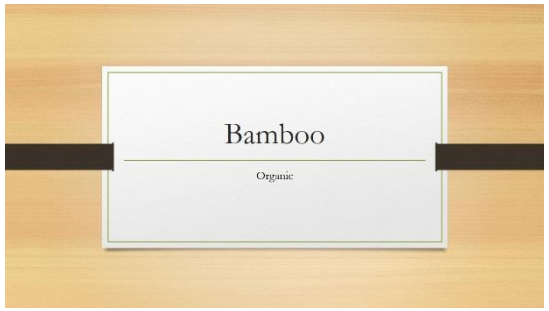
F1 score = 2 x (Precision x Recall) / (Precision + Recall)

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Significance Testing: Wilcoxon Signed-Rank Test

α=0.01



Machine Learning Models

Logistic Regression

C = 0.01, 0.1, 1, **10**, 100

k-Nearest Neighbor

No of features used: 8

k = **1**, 3, 5, 10

Decision Tree and Random Forest

No of estimators = 10, 50, 100, **200 (RF)**

Max leaf nodes = 32, 64, 128, **256 (RF/DT)**

Max features = 5, **10 (RF)**, **20 (DT)**

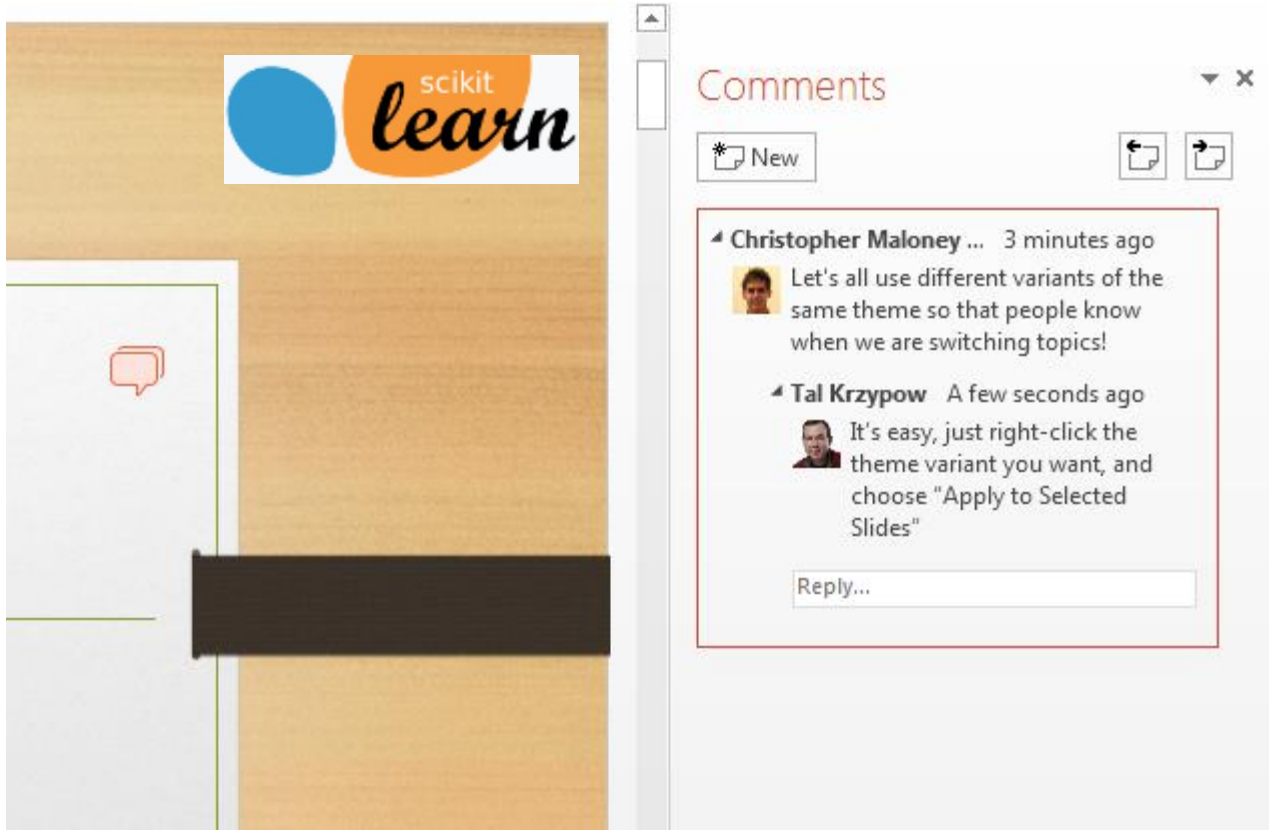
`sklearn.model_selection`.train_test_split

`sklearn.decomposition`.PCA

`sklearn.model_selection`.GridSearchCV

`sklearn.model_selection`.StratifiedKFold

`sklearn.metrics`.classification_report



Best parameters and scores

Model	Best Parameters	f1-score	Precision	Recall
Logistic Regression	C = 10	Class 0: 0.92 Class 1: 0.58	Class 0: 0.98 Class 1: 0.43	Class 0: 0.86 Class 1: 0.86
k-Nearest Neighbor	k = 1	Class 0: 0.93 Class 1: 0.53	Class 0: 0.96 Class 1: 0.45	Class 0: 0.90 Class 1: 0.66
Decision Tree	maximum leaf nodes = 256 maximum features = 20	Class 0: 0.94 Class 1: 0.64	Class 0: 0.98 Class 1: 0.51	Class 0: 0.90 Class 1: 0.85
Random Forest Classifier	No. of estimators = 200 Max leaf nodes = 256 Max features = 10	Class 0: 0.94 Class 1: 0.65	Class 0: 0.98 Class 1: 0.51	Class 0: 0.89 Class 1: 0.88

Logistic Regression				k-Nearest Neighbor				Decision Tree				Random Forest Classifier			
Prediction		No	Yes	**Prediction**		No	Yes	**Prediction**		No	Yes	**Prediction**		No	Yes
Actual	No	1,545	253	**Actual**	No	1,613	185	**Actual**	No	1,614	184	**Actual**	No	1,608	190
	Yes	31	194		Yes	76	149		Yes	34	191		Yes	27	198

The Wilcoxon test

All lower than α=0.01

	vs. Logistic Regression	vs. k-Nearest Neighbor	vs. Decision Tree
Random Forest	8.8574e-05	8.8574e-05	0.0001

The p-values from the Wilcoxon test when comparing the f1-scores

Q: "Is the recommended Random Forest model better than the alternatives considered?" **YES**

H_0: The recommended model's f1-score is not statistically better than the alternatives **REJECT**

H_a: The recommended model's f1-score is statistically better than the alternatives **ACCEPT**

What's Next

- Collect more data
- Use better features
- Use more complex models, eg neural network

→ Continuous Learning

Find out more, click here for further information

(Click the arrow for further information)