Hands on Machine Learning

Problem Statement - Titanic Disaster

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

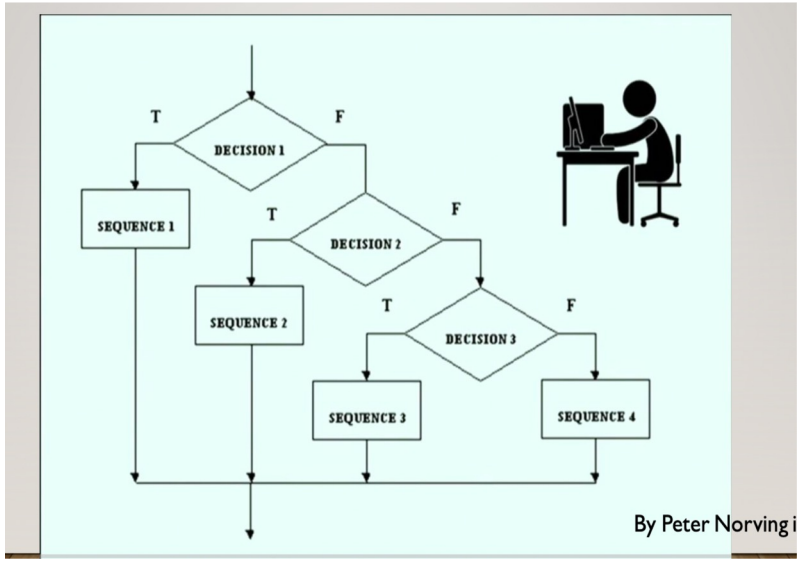
https://www.kaggle.com/competitions/titanic/overview/description

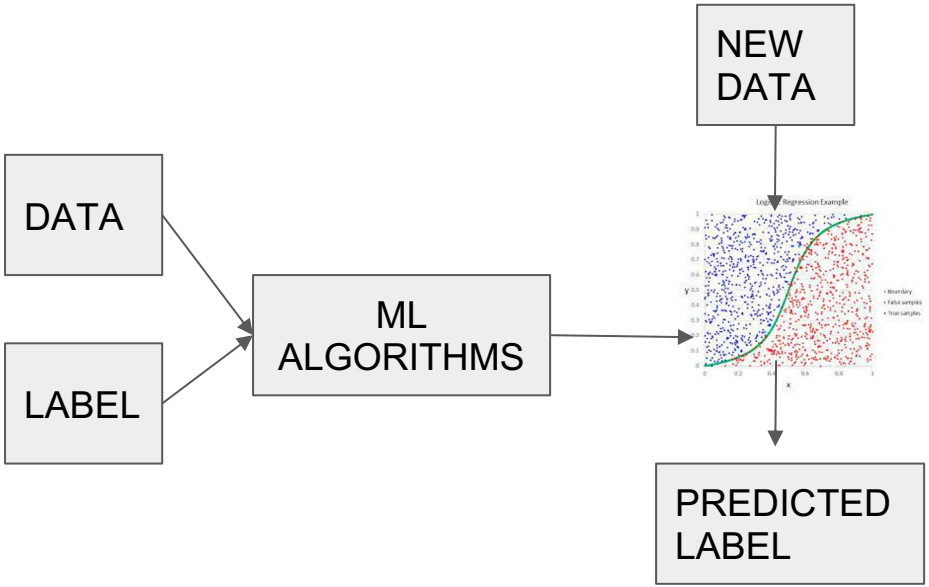
Data - Features

- `PassengerId` is the unique id of the row and it doesn't have any effect on target
- `Survived` is the target variable we are trying to predict (**0** or **1**):
  - **1 = Survived**
  - **0 = Not Survived**
- `Pclass` (Passenger Class) is the socio-economic status of the passenger and it is a categorical ordinal feature which has **3** unique values (**1**, **2** or **3**):
  - **1 = Upper Class**
  - **2 = Middle Class**
  - **3 = Lower Class**
- `Name`, `Sex` and `Age` are self-explanatory
- `SibSp` is the total number of the passengers' siblings and spouse
- `Parch` is the total number of the passengers' parents and children
- `Ticket` is the ticket number of the passenger
- `Fare` is the passenger fare
- `Cabin` is the cabin number of the passenger
- `Embarked` is port of embarkation and it is a categorical feature which has **3** unique values (**C**, **Q** or **S**):
  - **C = Cherbourg**
  - **Q = Queenstown**
  - **S = Southampton**

What is Machine Learning?

Traditional Programming - Rule based system





Supervised ML

Supervised Classification Problem

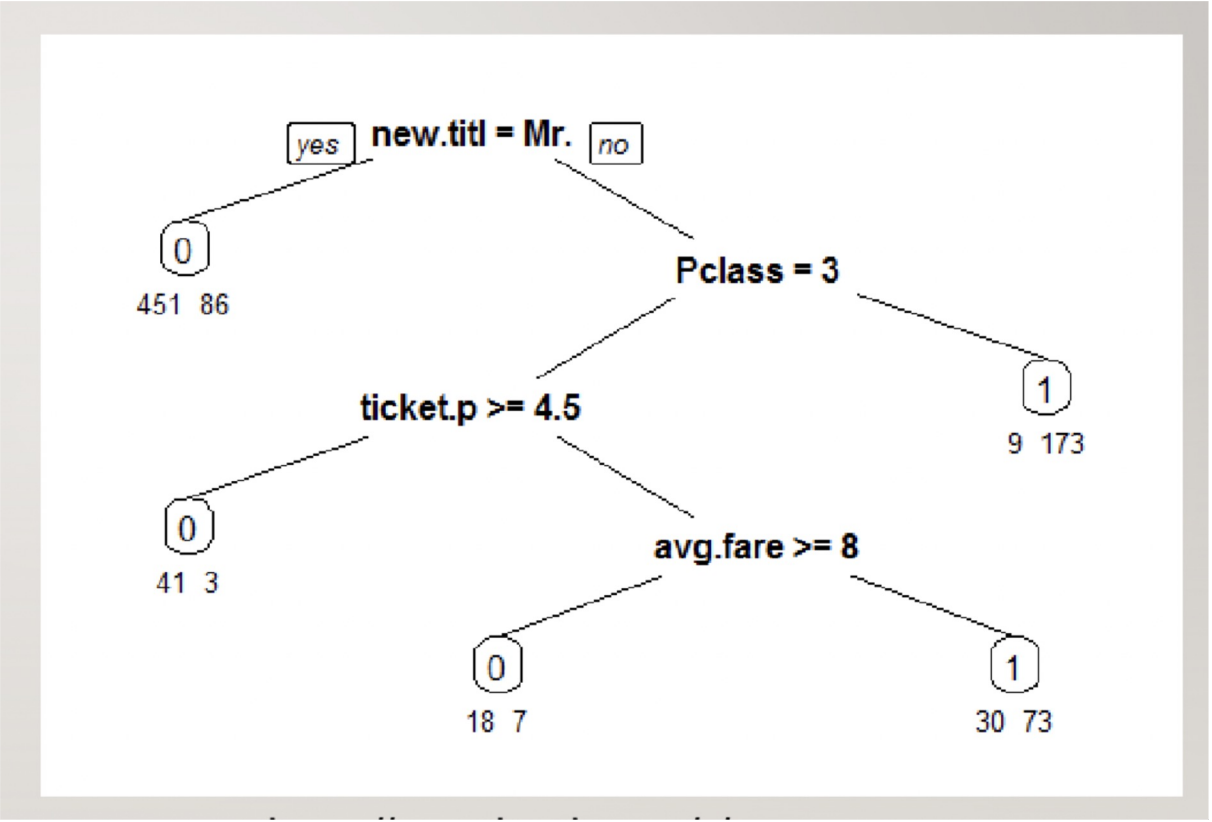
Output prediction is discrete classes (predict probabilities)

- Yes/No
- High/Medium/Low
- Cat/Dog
- Image Classification
- Product categorization

Supervised Regression Problem

Continuous values

- House Rent
- Loan amount



H2O installation

Install JRE

Download h20

http://h2o-release.s3.amazonaws.com/h2o/rel-zumbo/3/index.html

- cd ~/Downloads
- unzip h2o-3.36.1.3.zip
- cd h2o-3.36.1.3
- java -jar h2o.jar

http://localhost:54321

Metrics

For Classification

Accuracy – What % is correct (not suitable for imbalance dataset)

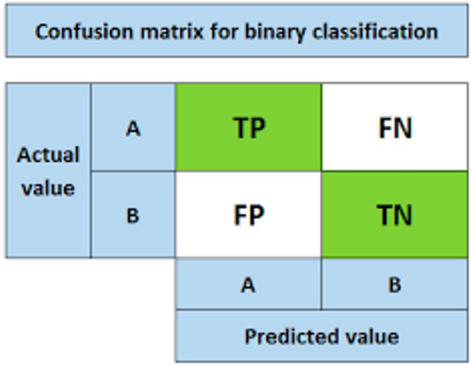
Precision - **Quality**
Of the ones predicted as True, what % is actually true.
TP/ (TP + FP)

Recall - **Quantity**
Of the ones which are actually positivem how much are predicted correctly
TP/ (TP + FN)

F-Score
Balance between both precision and recall.

For Regression

MAE, RMSE, MAPE, WMAPE, Bias %



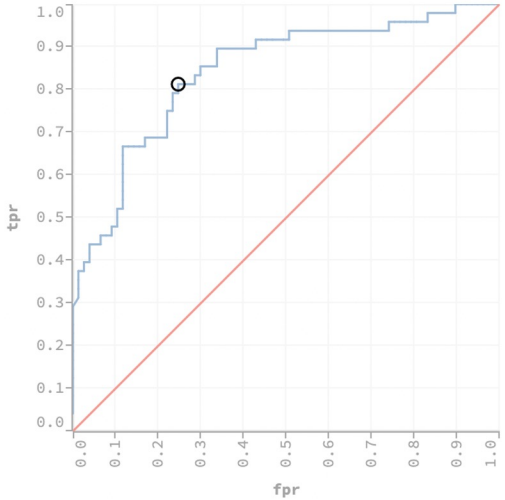
Questions:
Of 100 patients, 10 have cancer. Model predicts everyone has cancer? What is accuracy?

ML Model for Criminal justice - what is important - Precision or Recall?

Model to predict Covid or not - Precision or Recall?

First Model

▾ ROC CURVE - VALIDATION METRICS , AUC = 0.842938



Threshold: Choose...

Criterion: max f1

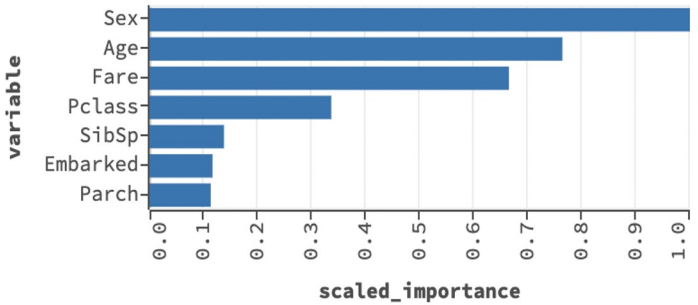
Selected mark(s):

threshold	0.3264
f1	0.7358
f2	0.7800
f0point5	0.6964
accuracy	0.7760
precision	0.6724
recall	0.8125
specificity	0.7532
absolute_mcc	0.5517
min_per_class_accuracy	0.7532
mean_per_class_accuracy	0.7829
tns	58
fns	9
fps	19
tps	39
tnr	0.7532
fnr	0.1875
fpr	0.2468
tpr	0.8125
idx	53

CM

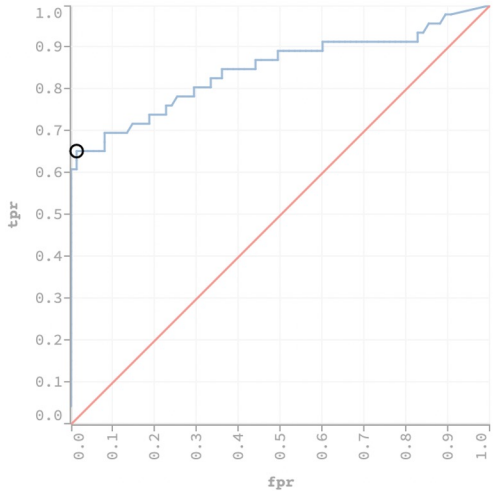
Actual/Predicted	0	1	Error	Rate
0	58	19	0.2468	19 / 77
1	9	39	0.1875	9 / 48
Total	67	58	0.2240	28 / 125

▾ VARIABLE IMPORTANCES



Feature Engineered Model

▾ ROC CURVE - VALIDATION METRICS , AUC = 0.847391



Threshold: Choose...

Criterion: max f1

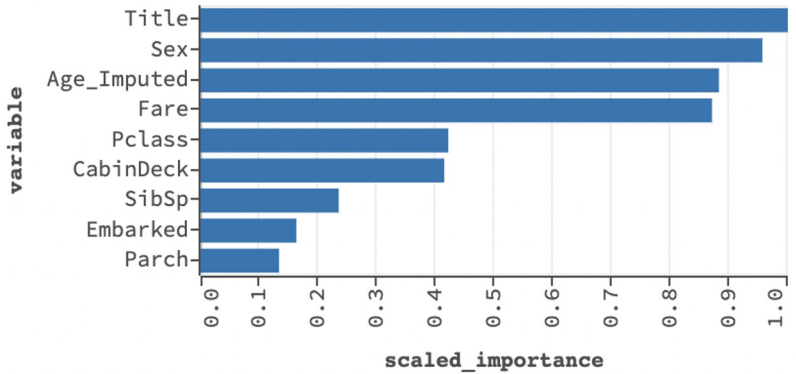
Selected mark(s):

threshold	0.7200
f1	0.7792
f2	0.6977
f0point5	0.8824
accuracy	0.8595
precision	0.9677
recall	0.6522
specificity	0.9867
absolute_mcc	0.7104
min_per_class_accuracy	0.6522
mean_per_class_accuracy	0.8194
tns	74
fns	16
fps	1
tps	30
tnr	0.9867
fnr	0.3478
fpr	0.0133
tpr	0.6522
idx	25

CM

Actual/Predicted	0	1	Error	Rate
0	74	1	0.0133	1 / 75
1	16	30	0.3478	16 / 46
Total	90	31	0.1405	17 / 121

▾ VARIABLE IMPORTANCES



Feature Engineered + Auto ML

Leaderboard

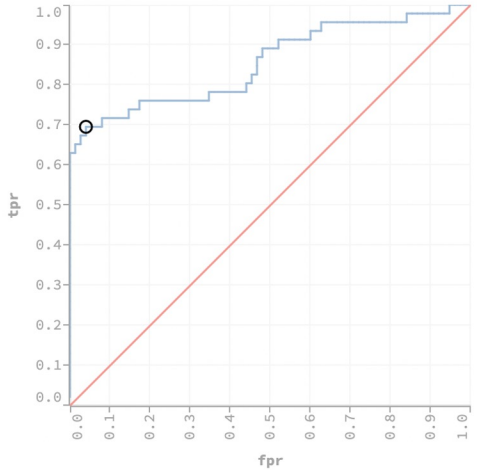
Monitor Live

MODELS

models sorted in order of auc, best first

	model_id	auc	logloss
0	StackedEnsemble_BestOfFamily_4_AutoML_1_20220727_233129	0.8850709887102292	0.4001344056388554
1	GBM_grid_1_AutoML_1_20220727_233129_model_2	0.8829969209716044	0.4095273200315345
2	GBM_grid_1_AutoML_1_20220727_233129_model_13	0.8817959573497549	0.407944612427401152
3	GBM_grid_1_AutoML_1_20220727_233129_model_10	0.8786385277682747	0.413183290114981
4	StackedEnsemble_AllModels_3_AutoML_1_20220727_233129	0.8785529992017334	0.4025558643777332
5	GBM_grid_1_AutoML_1_20220727_233129_model_15	0.8780077545900331	0.4166259193118573
6	GBM_grid_1_AutoML_1_20220727_233129_model_5	0.8777868057931348	0.4138634454404394
7	DeepLearning_grid_2_AutoML_1_20220727_233129_model_2	0.8770099213137188	0.4159123514786134
8	StackedEnsemble_AllModels_2_AutoML_1_20220727_233129	0.8769671570304483	0.4104476236659049
9	StackedEnsemble_BestOfFamily_3_AutoML_1_20220727_233129	0.8768317368000912	0.4132630228069975
10	StackedEnsemble_BestOfFamily_2_AutoML_1_20220727_233129	0.8760370338693124	0.412105176276588
11	GBM_grid_1_AutoML_1_20220727_233129_model_8	0.8760192154179496	0.4143006418211888
12	GBM_2_AutoML_1_20220727_233129	0.875666410080967	0.4163002294830256
13	DeepLearning_grid_1_AutoML_1_20220727_233129_model_19	0.8744654464591173	0.4467866598476236
14	DeepLearning_grid_2_AutoML_1_20220727_233129_model_16	0.8744155547953016	0.4227289080585543
15	DeepLearning_grid_1_AutoML_1_20220727_233129_model_3	0.8743870452731213	0.455560945253822
16	GBM_4_AutoML_1_20220727_233129	0.8740912589804996	0.4190816239568775
17	DeepLearning_grid_2_AutoML_1_20220727_233129_model_14	0.8735638328201619	0.4178616190289039
18	GBM_grid_1_AutoML_1_20220727_233129_model_16	0.8733713935454441	0.4155765678023712
19	StackedEnsemble_AllModels_1_AutoML_1_20220727_233129	0.8733464477135363	0.4157170508085113
20	DeepLearning_grid_2_AutoML_1_20220727_233129_model_8	0.8723771239594025	0.4289342626312643

ROC CURVE - VALIDATION METRICS , AUC = 0.855072



Threshold: Choose...

Criterion: max f1

Selected mark(s):

threshold	0.5963
f1	0.7901
f2	0.7306
f0point5	0.8602
accuracy	0.8595
precision	0.9143
recall	0.6957
specificity	0.9600
absolute_mcc	0.7019
min_per_class_accuracy	0.6957
mean_per_class_accuracy	0.8278
tns	72
fns	14
fps	3
tps	32
tnr	0.9600
fnr	0.3043
fpr	0.0400
tpr	0.6957
idx	34

CM

Actual/Predicted	0	1	Error	Rate
0	72	3	0.0400	3 / 75
1	14	32	0.3043	14 / 46
Total	86	35	0.1405	17 / 121

What is most important for ML Model?

- Data
  - Good quality data/label
  - EDA - Imputation, Outlier Detection
  - Feature Engineering
  - Domain Understanding (SMEs)
- Models
  - Hyper-parameter Tuning
  - Trying different Model types (Auto ML)
- Metrics
  - ML Metric
  - Business Metric
- Model Performance in real world
  - Model Performance Monitoring on ground truth

How could Jack survived with Rose?

Questions