WILD
CODE
SCHOOL

1912. Learning from disaster

Project Week 03

The sinking of the Titanic on April 15th in 1912 is still considered one of the biggest tragedies in naval history. It's also a tale of human hybris - since there were not nearly enough lifeboats to rescue all passengers and the crew.

Objective

Today, in this project, you have an extraordinary power: you can travel in time to try to save some passengers. You obviously wanted to save Jack and as many other people as possible. To do this, you have to identify the people who are most probably going to die.
The dataset that is provided has many different input columns and the target variable is "survived". It is your task to explore this data, fix it where necessary, encode, select, analyze, etc. and finally build an appropriate machine learning classification model to predict if a passenger would survive or die and get the highest possible accuracy score.

The Challenge

This project was part of a Kaggle Challenge.
You can find the test and train data set as well as descriptions of the features there.

1. Data Wrangling & Cleaning (look out for missing values)
3. Data Exploration (also: How many people named Jack have been on board?)
4. Feature Engineering
5. Train-Validation-Split (using train_test_split of sklearn)
6. Train the model
7. Validate the model (accuracy as well as confusion matrix)
8. Include two more algorithms and compare the results
9. Load your results on kaggle and check your score

Bonus: Find the 100 people in the training set with the highest probability of dying. How many actually survived?

Technical Reference

Feature Engineering

- OneHotEncoding
- Binning
- Scaling

Classifiers in Sklearn

- Naive Bayes
- kNN - k Nearest Neighbors
- SVM - Support Vector Machines
- Logistic Regression
- Random Forest

Technical Reference

k-fold Cross Validation

Confusion Matrix

Deliverables

A clean jupyter notebook with comments on the code and a clear structure.
It should present and describe each step of the process and contain a variety of plots for the data exploration and your findings as well as comparisons of the algorithms.

You present your jupyter notebook on friday - additional slides are not needed!

Good Luck with your work!