Assignment 4: Kaggle Competitions

June 28, 2022

1 A history of Kaggle - from The Kaggle Book

Kaggle took its first steps in February 2010, thanks to Anthony Goldbloom, an Australian trained economist with a degree in Economics and Econometrics. After working at Australia's Department of the Treasury and the Research department at the Reserve Bank of Australia, Goldbloom interned in London at The Economist, the international weekly newspaper on current affairs, international business, politics, and technology. At The Economist, he had occasion to write an article about big data, which inspired his idea to build a competition platform that could crowdsource the best analytical experts to solve interesting machine learning problems https://www.smh.com.au/technology/from-bondi-to-the-big-bucks-the-28yearold-whos-making-data-science-a-sport-20111104-1myq1.html. Since the crowdsourcing dynamics played a relevant part in the business idea for this platform, he derived the name Kaggle, which recalls by rhyme the term gaggle, a flock of geese, the goose also being the symbol of the platform.

After moving to Silicon Valley in the USA, his Kaggle start-up received 11.25 USD million in Series A funding from a round led by Khosla Ventures and Index Ventures, two renowned venture capital firms. The first competitions were rolled out, the community grew, and some of the initial competitors came to be quite prominent, such as Jeremy Howard, the Australian data scientist and entrepreneur, who, after winning a couple of competitions on Kaggle, became the President and Chief Scientist of the company.Jeremy Howard left his position as President in December 2013 and established a new start-up, fast.ai , offering machine learning courses and a deep learning library for coders.

At the time, there were some other prominent Kagglers (the name indicating frequent participants of competitions held by Kaggle) such as Jeremy Achin and Thomas de Godoy. After reaching the top 20 global rankings on the platform, they promptly decided to retire and to found their own company, DataRobot. Soon after, they started hiring their employees from among the best participants in the Kaggle competitions in order to instill the best machine learning knowledge and practices into the software they were developing. Today, DataRobot is one of the leading companies in developing AutoML solutions (software for automatic machine learning).

The Kaggle competitions claimed more and more attention from a growing audience. Even Geoffrey Hinton, the "godfather" of deep learning, participated in (and won) a Kaggle competition hosted by Merck in 2012 https://www.kaggle.com/c/MerckActivity/overview/winners. Kaggle was also the platform where François Chollet launched his deep learning package Keras during the Otto Group Product Classification Challenge https://www.kaggle.com/c/otto-group-product-classification-challenge/discussion/13632 and Tianqi Chen launched XGBoost, a speedier and more accurate version of gradient boosting machines, in the Higgs Boson Machine Learning Challenge https://www.kaggle.com/c/higgs-boson/discussion/10335.

2 Ensembles, Gradient Boosting and Neural Networks - 8 Points

Using and Ensemble Method, a Gradient Boosting Method and a Neural Network Method, provide a solution to each of the following famous Kaggle Competition:

1. The Higgs Boson Machine Learning Challenge
2. The Kepler Exoplanets Survey
3. The Ames House Advanced Regression dataset
4. The Lendning Club Dataset

Compare the solutions in each of the cases.

3 The LightGBM Package - 2 Points

Microsoft has it's on implementation of Gradient Boosting, the LightGBM. It is supposed to be faster than others implementations and more or as precise as it's competitors, such as XGBoost and CatBoost. In class, we evaluated the Credit Risk dataset using both XGBoost and CatBoost. Now, provide a solution using LightGBM, that must contain a baseline model and an optimised model for the following problems:

1. The Car Price Dataset
2. The Credit Risk Dataset

Compare both the baseline and optimised solution with it's pairs using XGBoost and CatBoost.