Table of Contents

[Abstract 1](#_Toc493500397)

[Introduction 1](#_Toc493500398)

[Two Classification Problems Considered 1](#_Toc493500399)

[Training and Test Results 1](#_Toc493500400)

[Analysis of models 1](#_Toc493500401)

# Abstract

There is no free lunch in data science. Flesh this out more.

# Introduction

We implemented the following five learning algorithms:

* Decision tree with pre-pruning
* Neural network
* Gradient Boosting
* Support Vector Machine
* K Nearest Neighbors

In addition, we introduced two separate classification problems as separate platforms to deploy the learning algorithms. Training and testing errors using each of the learning algorithms are then recorded and analyzed.

# Two Classification Problems Considered

1. Freddie Mac mortgages loan-level performance data

Fannie Mae and Freddie Mac began reporting loan-level credit performance data in 2013 at the direction of their regulator, the Federal Housing Finance Agency. The stated purpose of releasing the data was to “increase transparency, which helps investors build more accurate credit performance models in support of potential risk-sharing initiatives.”

Freddie Mac, one of the government-sponsored enterprises went through a nearly [$200 billion government bailout](http://en.wikipedia.org/wiki/Federal_takeover_of_Fannie_Mae_and_Freddie_Mac) during the financial crisis, motivated in large part by losses on loans that they guaranteed, so I figured there must be something interesting in the loan-level data.

I decided to dig into this dataset – to analyze multiple-period mortgage risk at loan level using the Freddie Mac dataset prime and subprime mortgages originated in the United States in 2016, which includes the individual characteristics of each loan, and monthly updates on loan performances over life of a loan.

This problem is interesting because there exists a highly nonlinear relationship between the variables and the default prediction. In particular, many of the existing academic research studied deep neural networks, which have multiple layers of hidden nodes. In this dataset, I am more interested in finding out which of the 5 classifiers work the best (without resorting to deep neural networks):

* Decision tree with pre-pruning
* Neural network
* Boosting
* SVM
* KNN

This is particularly feasible given the large size of my dataset (> 200,000 mortgages).

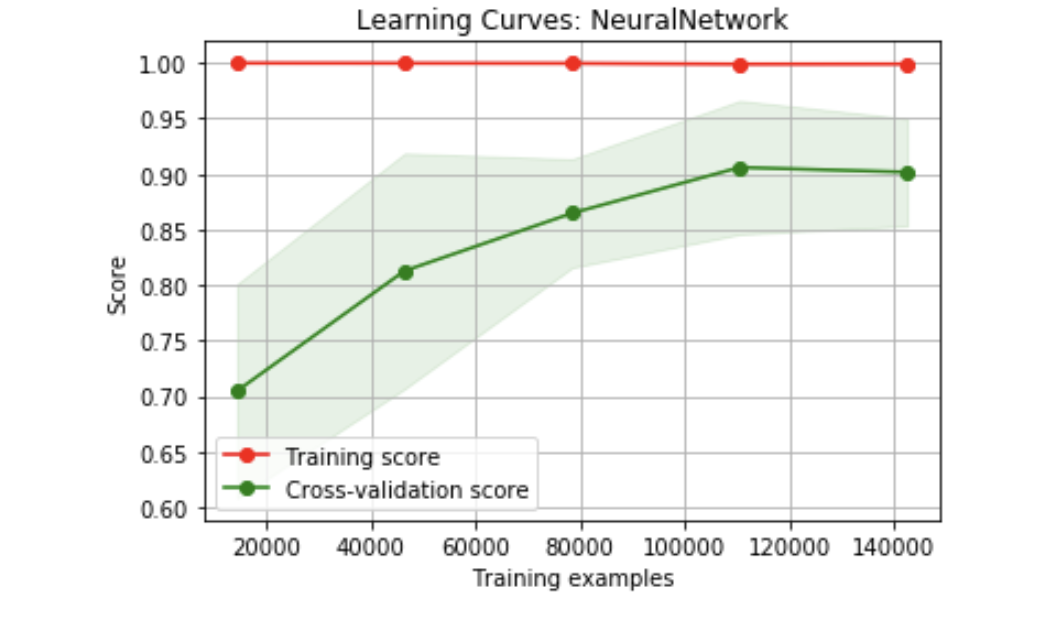
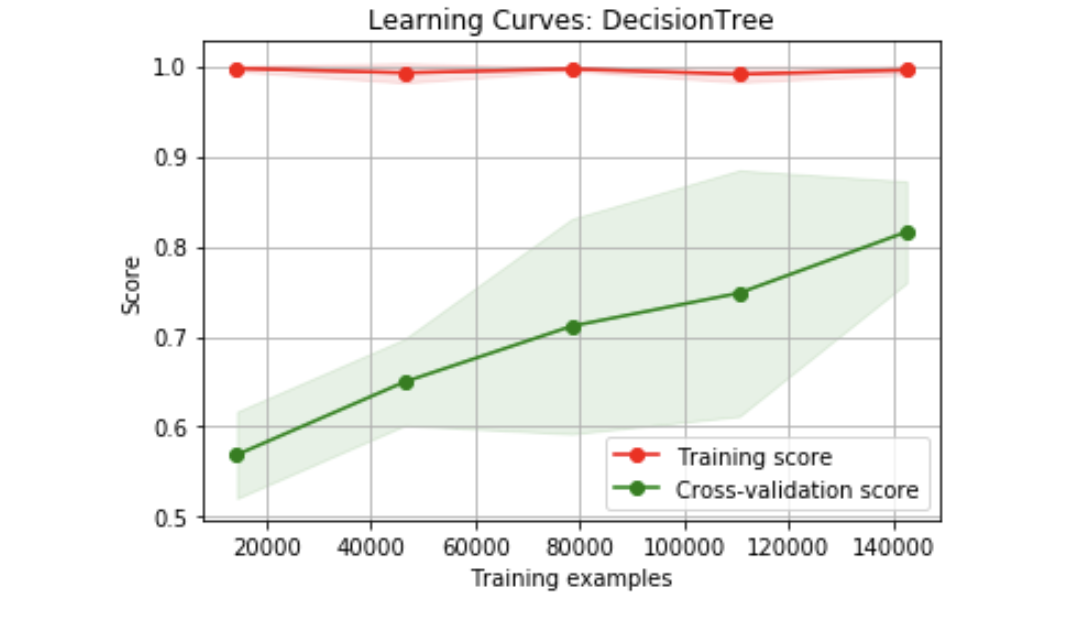
1. Blood donation prediction

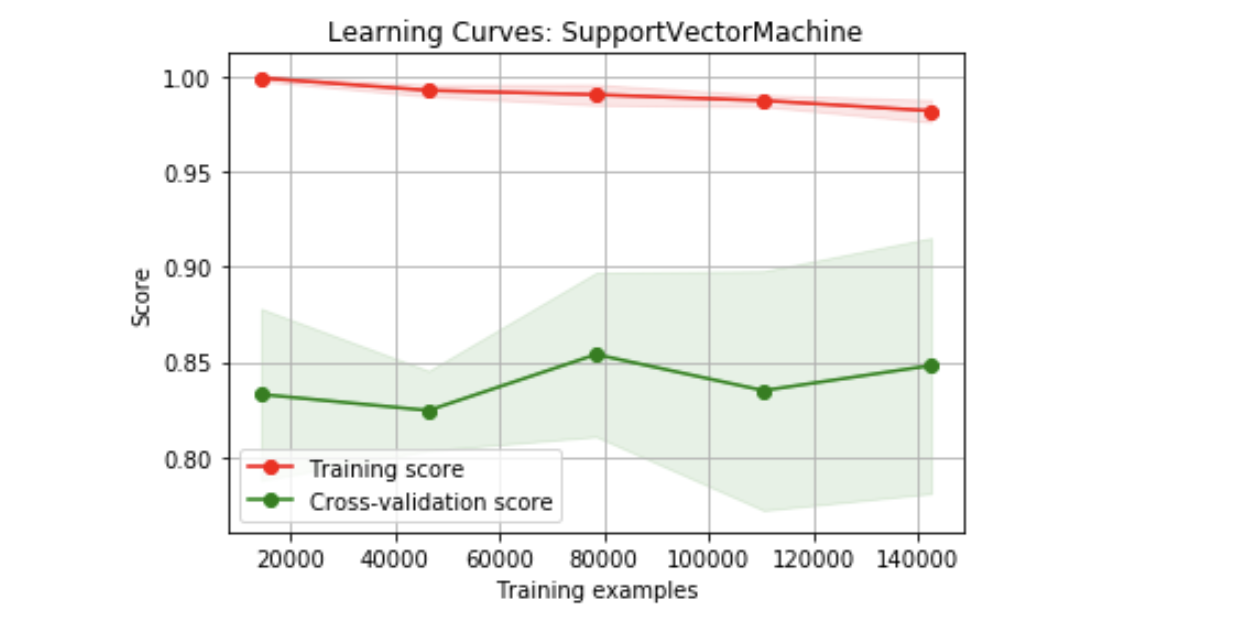
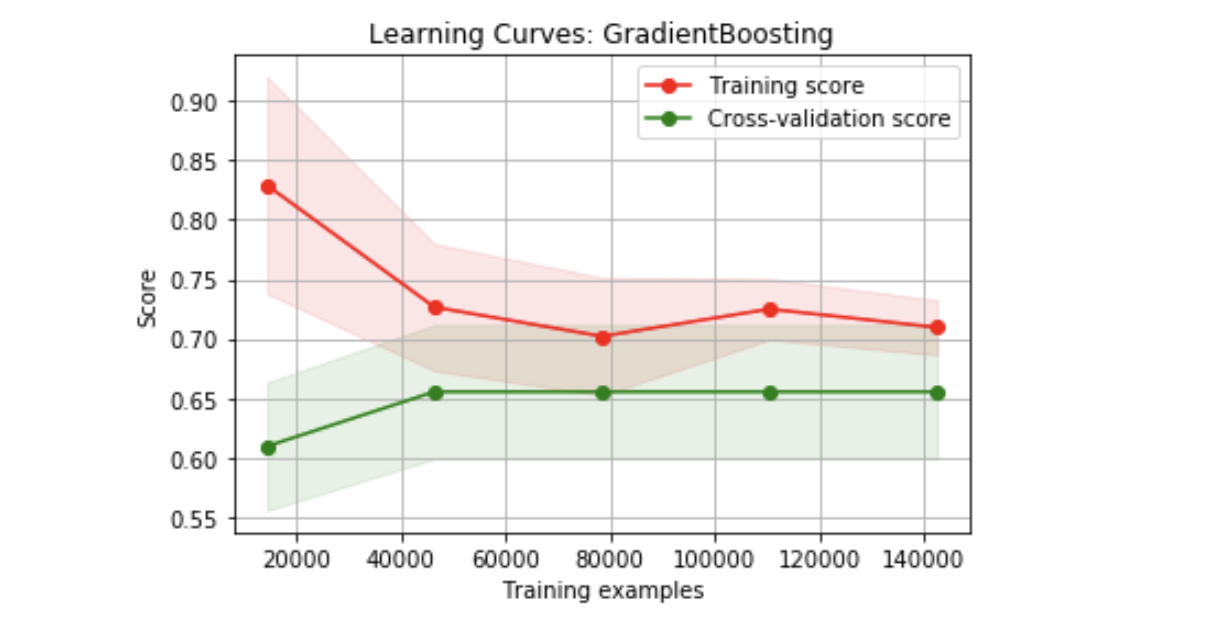
The [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.html) is a great repository of data science-related projects. Given my interest in discovery our mission, we're interested in predicting if a blood donor will donate within a given time window.

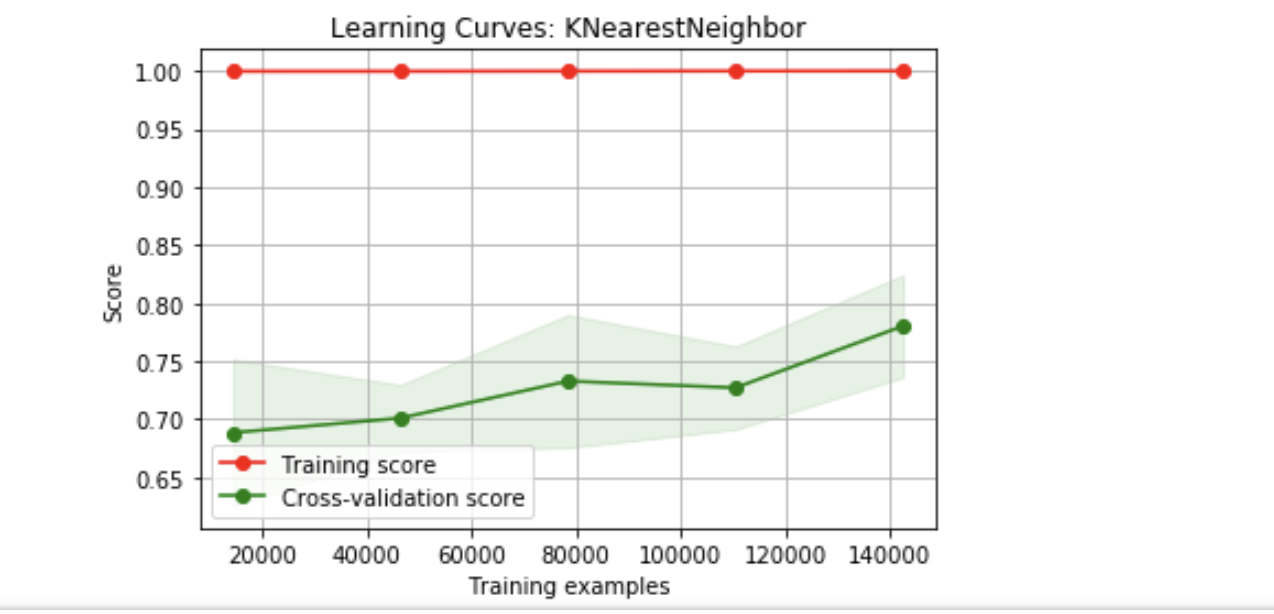
# Training and Test Results

1. Freddie Mac loan-level dataset

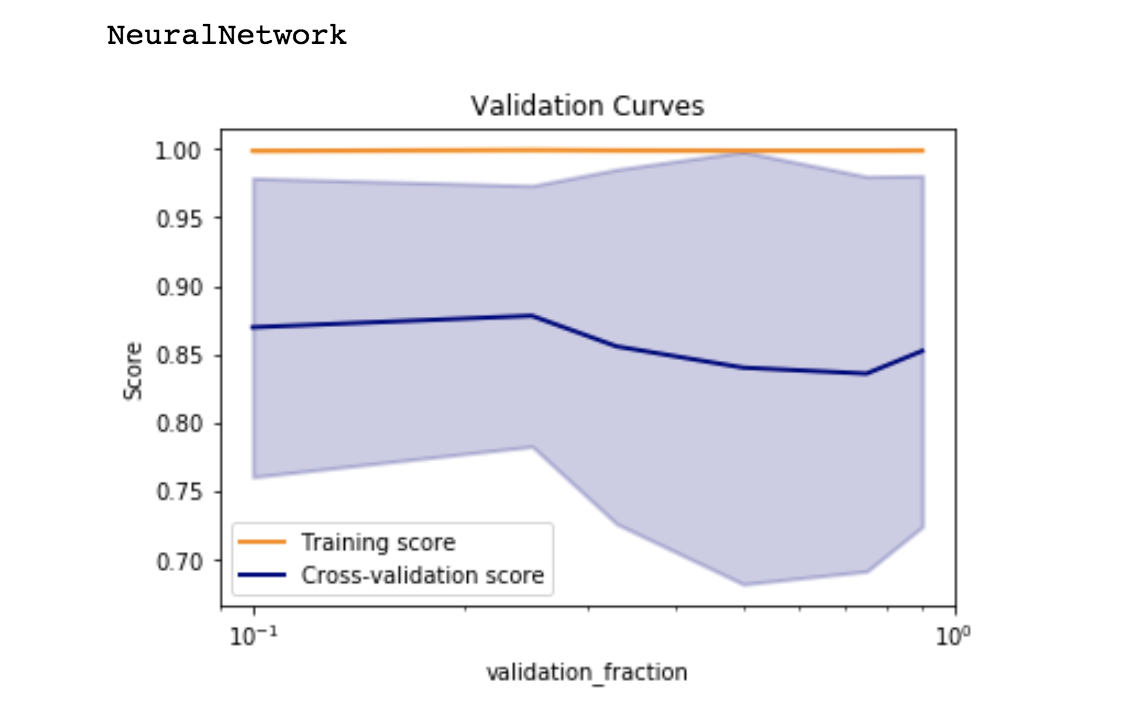
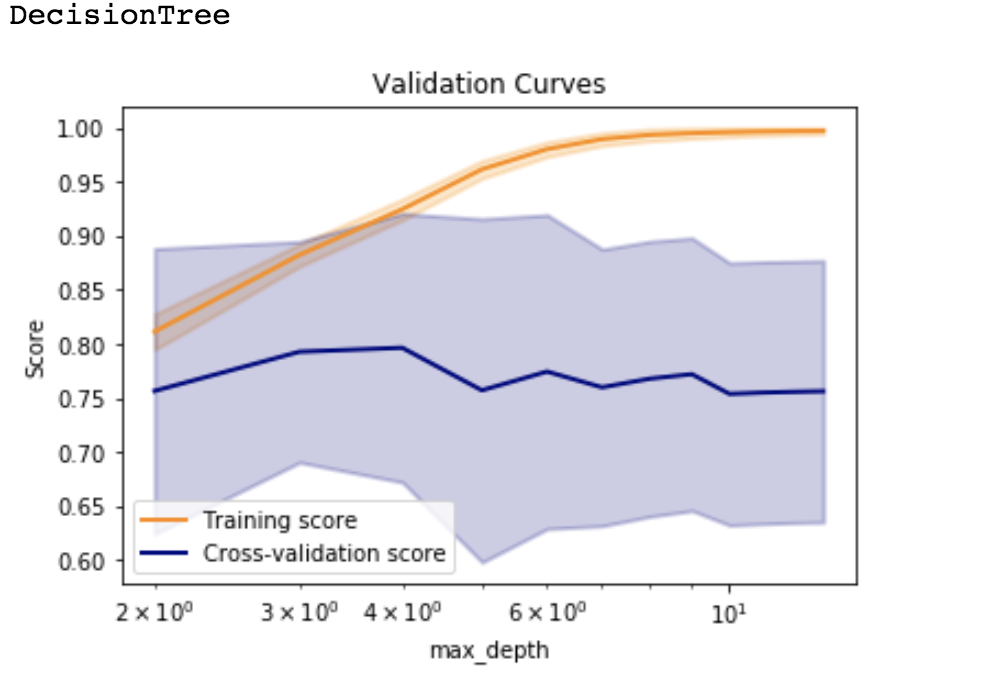
Learning curve analysis

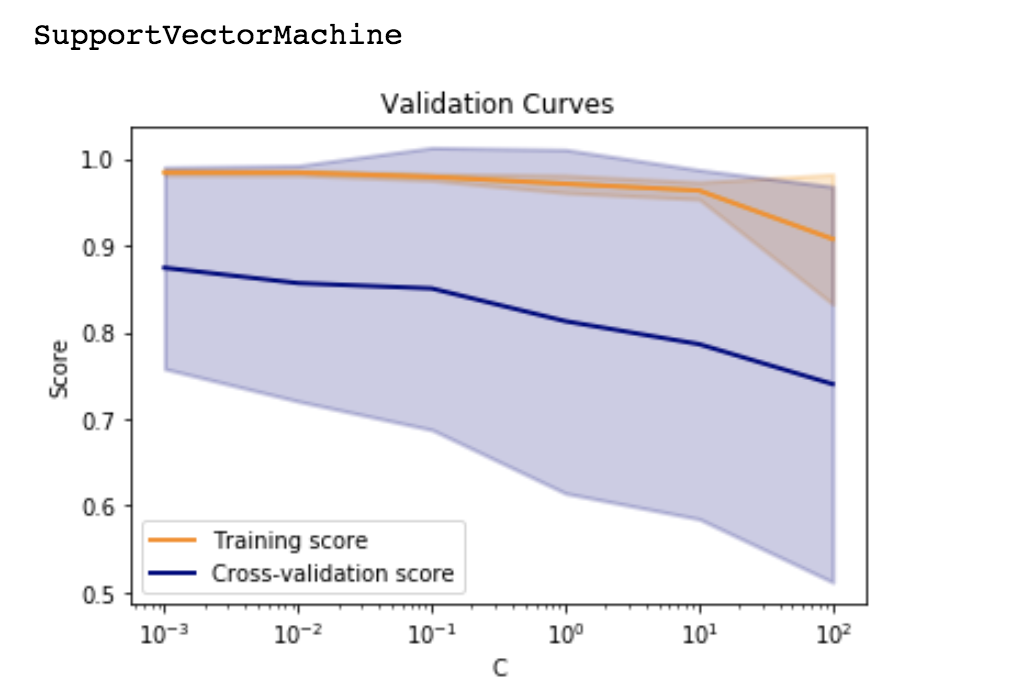
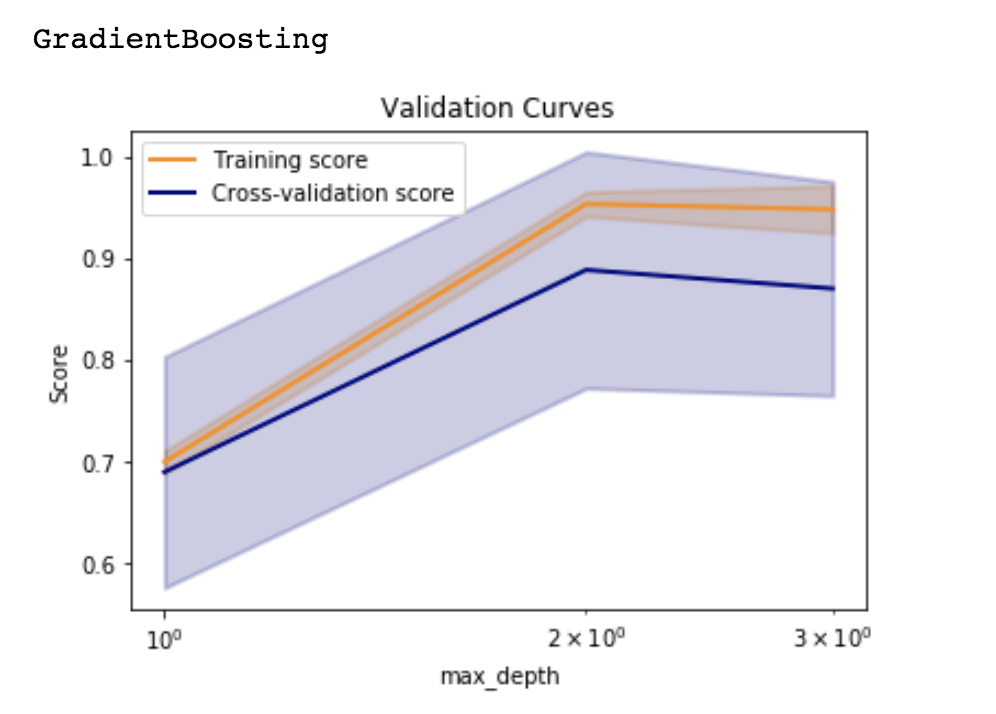


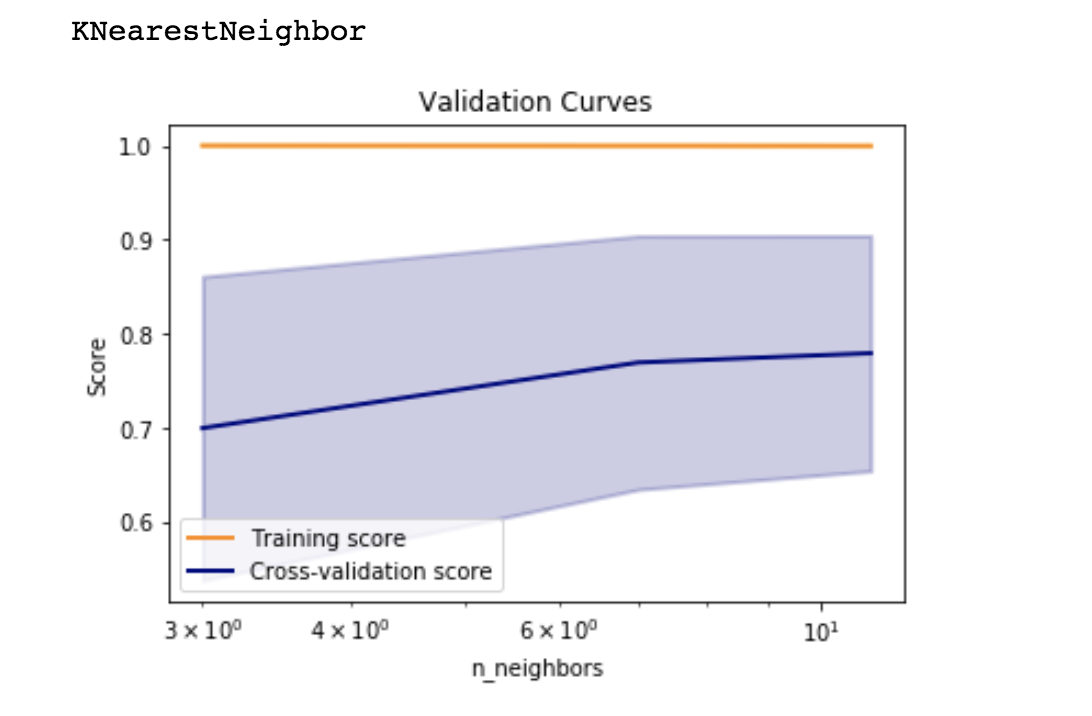




Complexity analysis

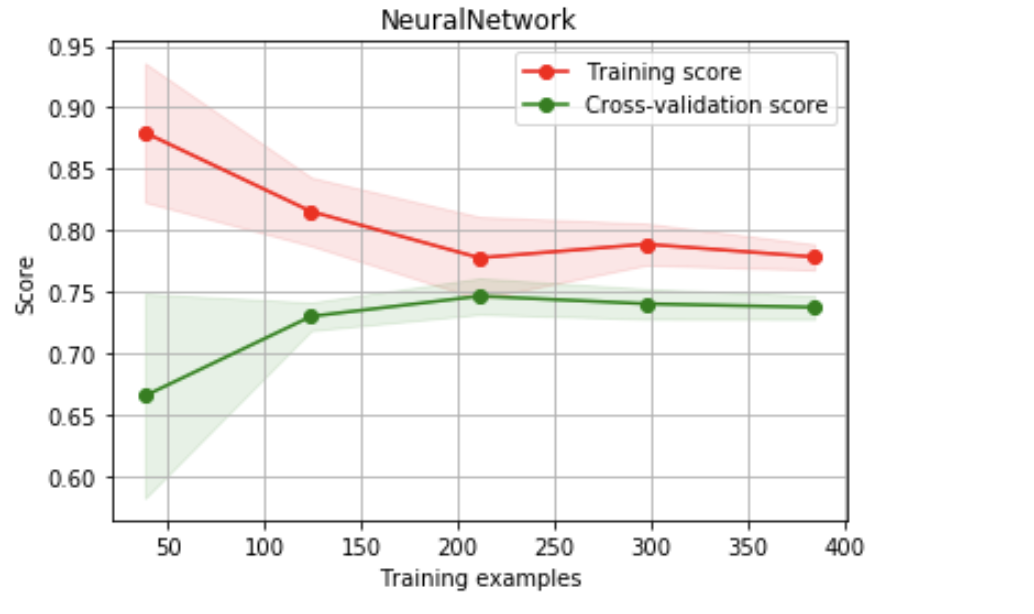
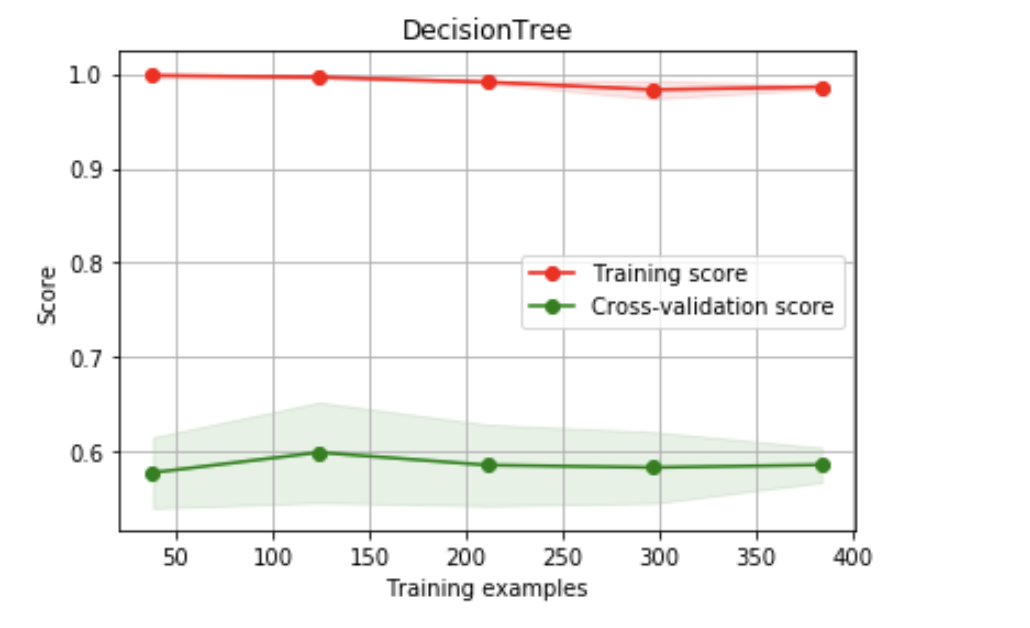


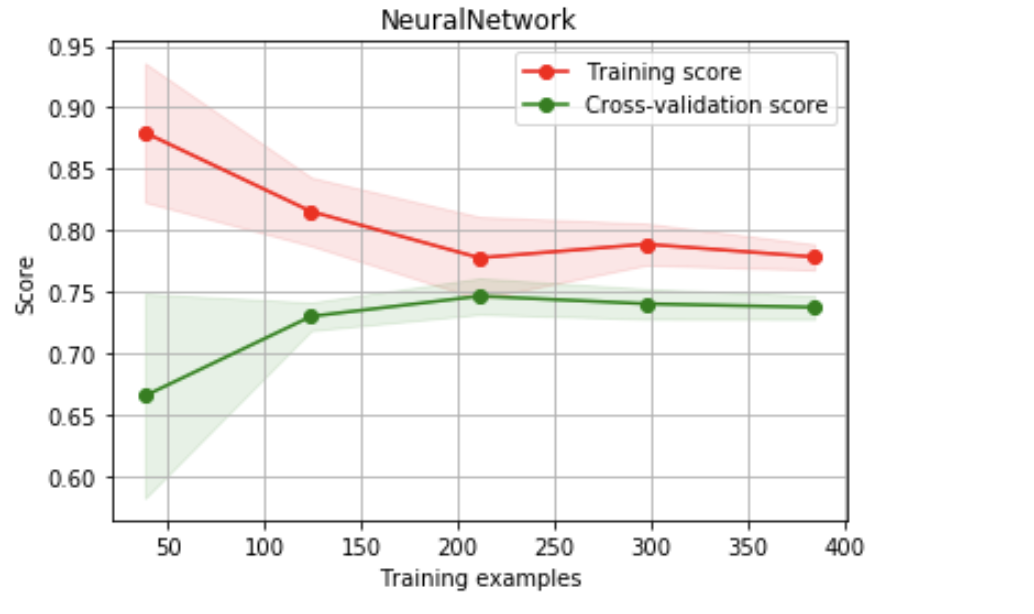
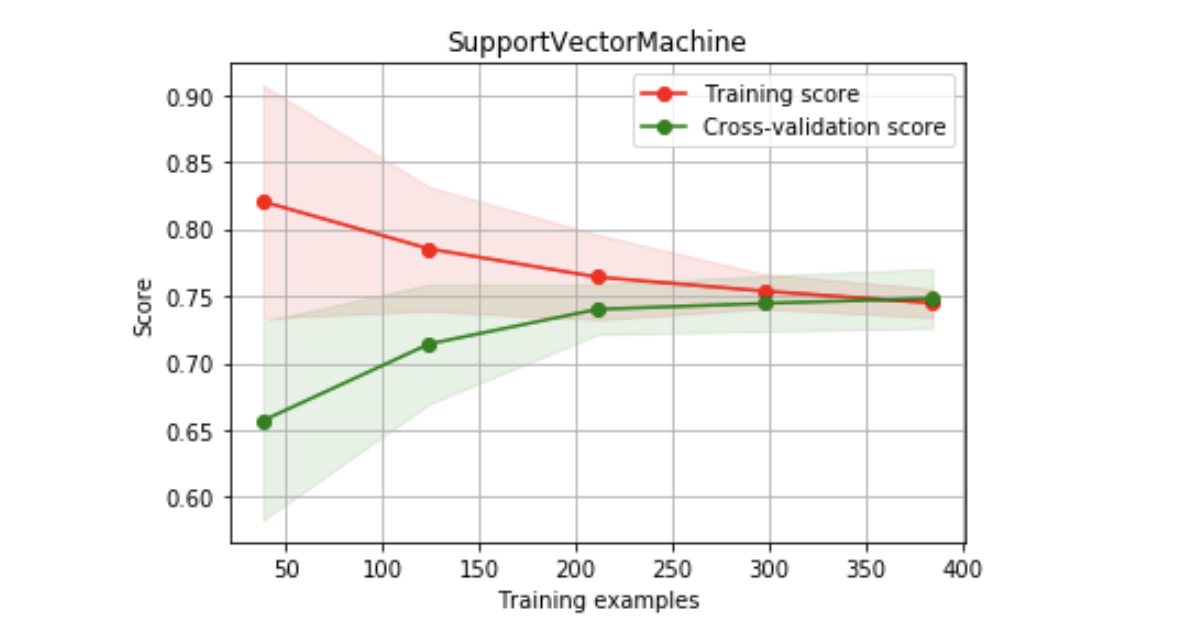


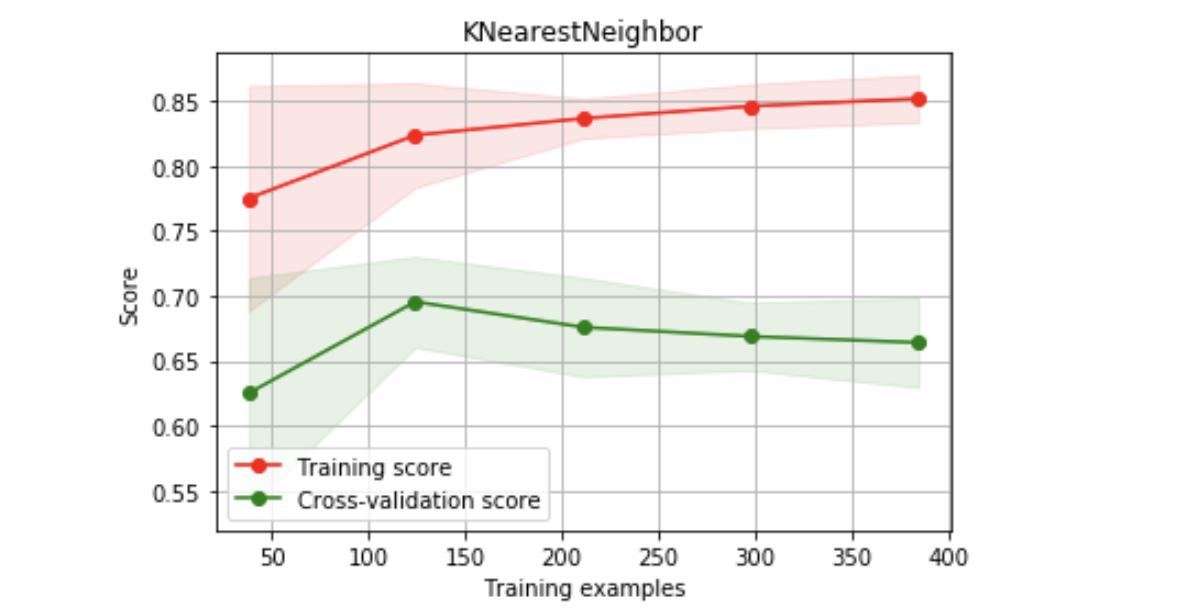


1. Blood donation prediction dataset

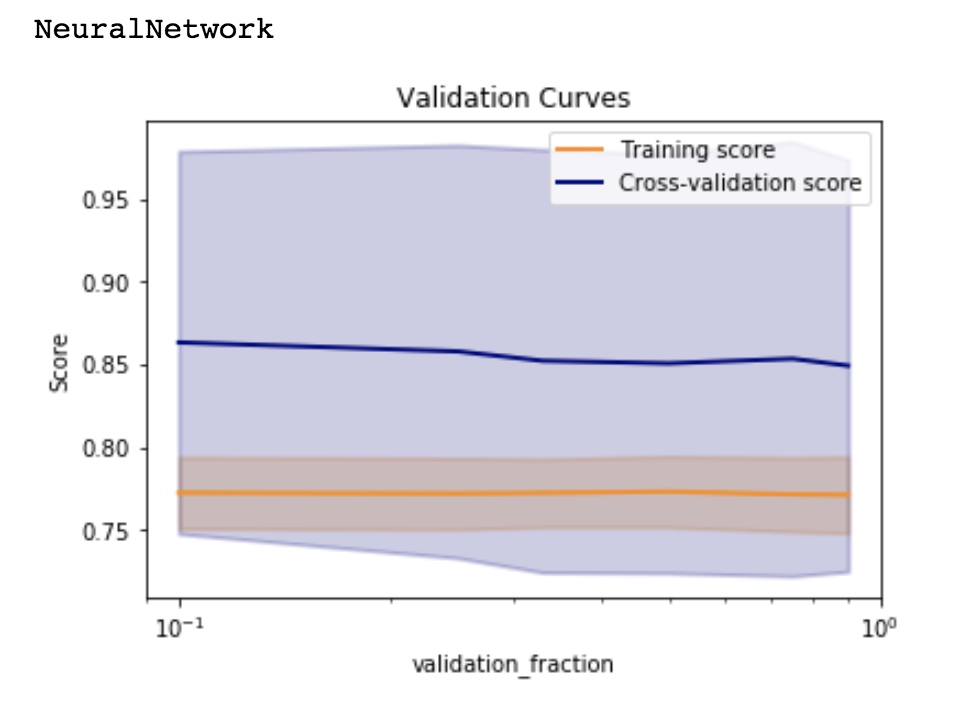
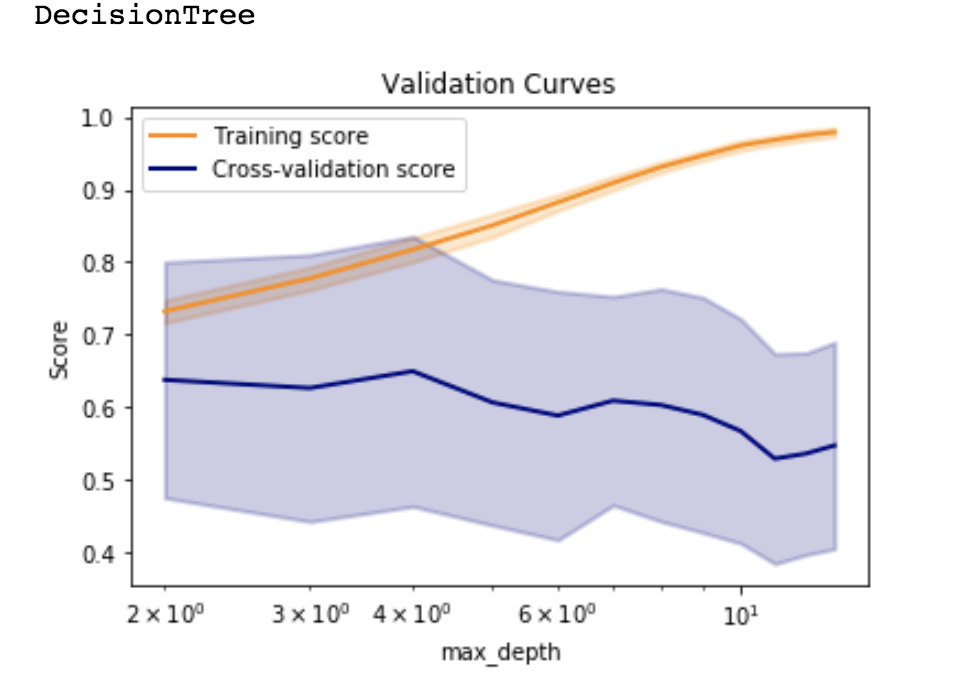
Learning curve analysis

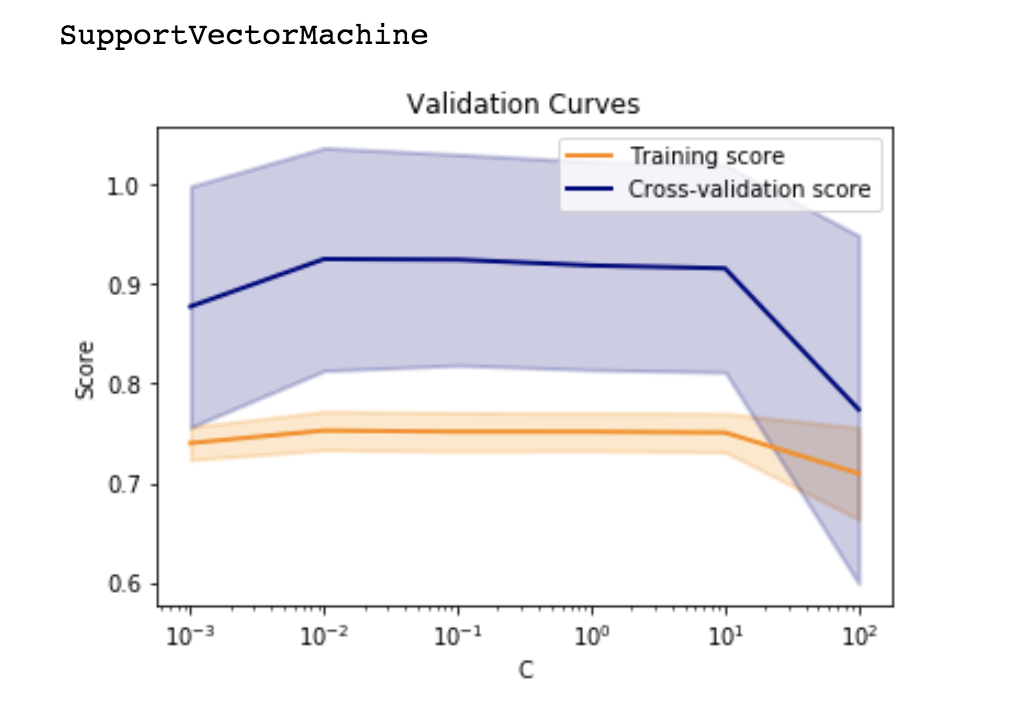
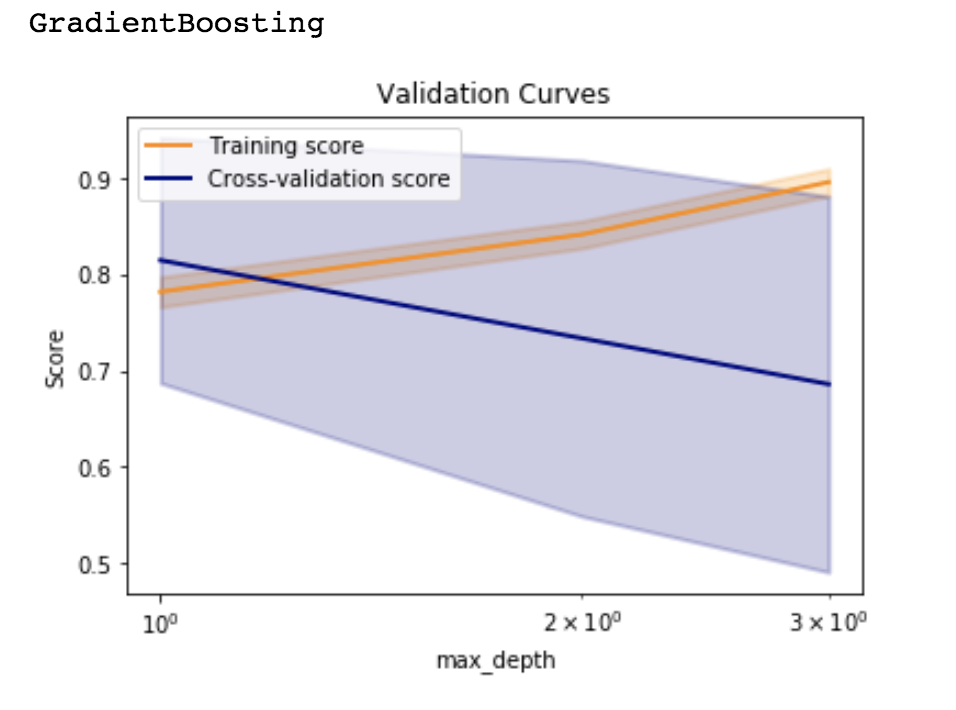


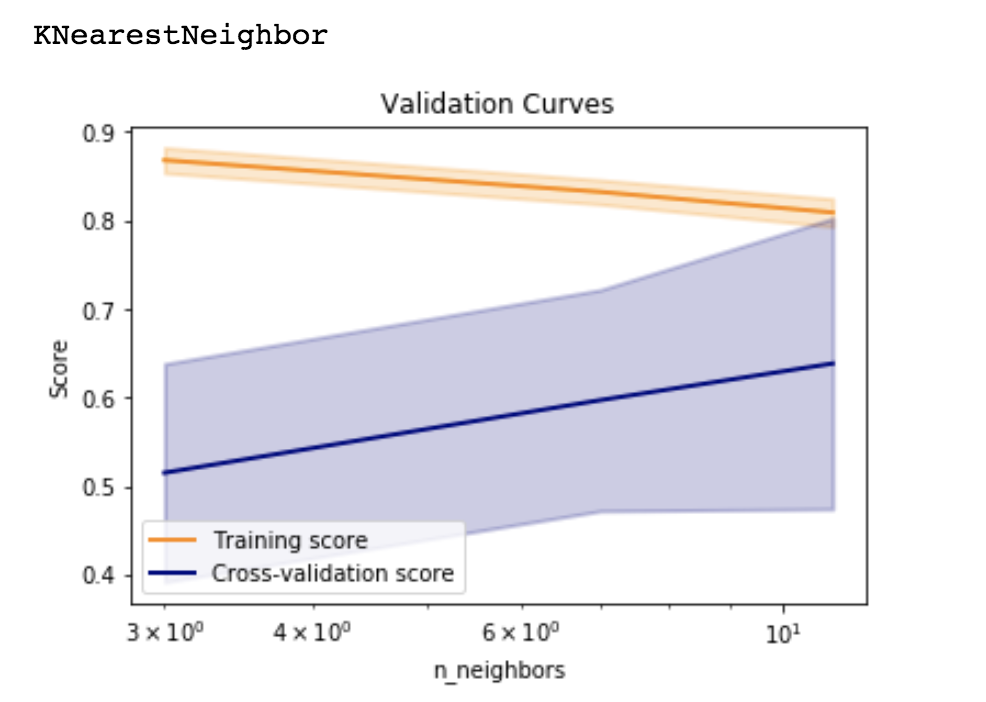




Complexity analysis







# Analysis of models