CE 475
Fundamentals and Applications of Machine Learning

KEREM CAN ÖZCAN
20140601039

Identification and Significance of the Problem

In this project, I tried to figure out missing values from the given dataset from the instructor of this course. The aim of the project is to predict the output for the given input label. I used sci-kit learn library to implement the methods that we covered in the class. And, trying to predict something with the given information is a great thing, it is definitely worth to try.

Methodology

First of all, I did research for what I should use in this program, what is the best ways of doing it. I decided to use Multiple Linear Regression, Polynomial Regression, Support Vector Machines, Decision Tree, and Random Forest.

Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory variables and the response variable.

Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.

Support Vector Machines

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Random Forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Implementation

I used the Python programming language. In my computers I had Spyder as an Integrated Development Environment. I used the libraries that;

- Numpy
- Pandas
- Matplotlib.pyplot
- Sklearn.linear_model
- Sklearn.preprocessing
- Sklearn.tree
- Sklearn.ensemble
- Sklearn.svm

Results

Backward Stepwise Elimination 10-fold Cross-Validation:

Mean = -64.015
RMSE = 789.192

Polynomial Regression 10-fold Cross-Validation:

Mean = -226562.062
RMSE = 28607.612

Support Vector Machine 3-fold Cross-Validation:

Mean of SVC = 0.145
Mean of SVR = -0.296
RMSE = 694.774

Decision Tree 10-fold Cross-Validation:

Mean = 0.693
RMSE = 374.102

Random Forest 1000 10-fold Cross-Validation:

Mean = 0.789
RMSE = 265.891

*These values above are not the best solutions.

Conclusion

These are the scatter plots of our project.X2 and X6 look the same. I tried Multiple Linear Regression, Backward Stepwise Elimination, Polynomial Regression, Decision Tree, Random Forest, Support Vector Machines respectively. I got the best results from Random Forests with 1000 trees. It could give better results according to the number of trees. When I did not use the '*random_state*' parameter I got cross-validation values between 75% to almost 90% for random forest regression. I cannot tell if one can success to extract the formula of those given table by implementing regression, but with more training data, we will be able to complete the task more successfully. The rest lies in my project. I tried to learn and gain lots of different things in machine learning and their methods. Having a chance of learning about machine learning methods was very exciting. Finally, I should be learning more and more about these methods and Python in the close future.