School of Computer Science
The University of Adelaide

Concepts in Artificial Intelligence & Machine Learning
Assignment 2

Trimester 1, 2022
Due 11.59 pm, Sunday 20 March 2022

1. Introduction

The goal of assignment 2 is to help you understand the basic flow of machine learning. The task is to predict the sales price for each house by building an advanced regression model. You need to write Python code to predict the sales price and analyze the impact of different factors based on your model. The total score is 30 marks. You are required to submit (1) the runnable codes (22 marks) and (2) a report in PDF (8 marks) that includes the methodology, analysis and results. Please **submit a .zip file via MyUni. Each part has a hurdle,** which means you need to get at least 11 marks for (1) **and** at least 4 marks for (2) to get pass for this assignment.

2. Tasks

Your task is to build a regression model to predict the sales price for each house. The provided dataset includes 38 features of the houses and their corresponding house prices (Please refer to section 3 for a detailed description). You are free to choose the regression models. Specifically, the tasks you need to complete are:

(1) There are missing values in the provided dataset for both training and test set. Will the missing values affect the model training? How do you solve the issue? Do research on the possible solutions, choose two solutions and justify the reason you choose them. Write your own code to apply the two solutions and compare the regression model's performances on the datasets that apply the two solutions. Section 5 gives reference about handling missing values. (6 Marks)

(2) Write Python code to build a regression model from the training set that you applied the better solution in (1). Write the code to evaluate its performance on the test set that you applied the better solution in (1) with Root-Mean-Squared-Error (RMSE). RMSE shares similar motivation of the least square loss we taught in the lecture. Section 5 gives reference for the definition. Please try at least two regression models and compare their performances. (6 Marks)

(3) There are 38 features provided for each house. Are all of the features used in your regression models? What's the impact if removing some features? Please write your own code to compare different feature selections and justify your selection. At least two feature selections are investigated (full feature set, a subset of the features). The dataset is the modified ones that you applied the better solution in (1). (6 Marks)

3. Data Description

(1) train.csv Has 40 columns, the first column is the house ID, and the following 38 columns are features. The labels are in the last column SalePrice. There are three features that have missing value problems: OverallQual, OverallCond, YearBuilt.

Data Example:

ID	MSSubClass	LotArea	......	YrSold	SalePrice
1	60	8450	......	2008	208500

2	20	9600	……	2007	181500

(2) test.csv   Has 40 columns, the first column is the house ID, and the following 38 columns are features. The labels are in the last column SalePrice. Same as the training set, there are three features that have missing value problems: OverallQual, OverallCond, YearBuilt. The format of test.csv is the same as train.csv.

(3) data_description.txt - Detailed description of each column.

4. Deliverables

(1) Runnable codes in Python. (task: 18 marks, instructions and comments: 4 marks)

The codes can be .py files or a jupyter notebook file with clear instructions including required packages, how to run the codes (if applicable). You need to write comments for your code for comprehension and also as the best practices. Required packages must be listed in the requirement.txt file.

(2) A report in PDF format (8 marks) should have the following sections:

a. A description of how to handle the missing values in your code and report the results. (2.5 Marks)

b. A description of the regression technique you used and report the results. (3 Marks)

c. A description of the feature selection you applied and report the results. (2.5 Marks)

Note: Please use visualizations (figures or tables) to report the results and provide descriptions of the results.

5. Useful Resources

(1) Root-Mean-Squared-Error (RMSE)
https://en.wikipedia.org/wiki/Root-mean-square_deviation

(2) Guide to code commenting
https://www.codeconquest.com/advanced-programming-concepts/code-commenting/

(3) Guide to missing values
https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/