Assignment- Advanced Regression

**Problem Statement - Part I**

This assignment contains two parts. Part-1 is a programming assignment (to be submitted in a Jupyter notebook) whereas part-2 includes subjective questions (to be submitted in a PDF file).

Part-2 is given on the next page.

**Assignment Part-I**

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual value and flip them at a higher price. For the same purpose, the company has collected a data set from house sales in Australia. The data is provided in the csv file below.

The company is looking at prospective properties to buy to enter the market.

You are required to build a regression model using regularization, so as to predict the actual value of the prospective properties and decide whether to invest in them or not.

The company wants to know:

* Which variables are significant in predicting the price of a house
* How well those variables describe the price of a house

Also, determine the optimal value of lambda for ridge and lasso regression.

**Business Goal**

You are required to model the price of houses with the available independent variables. It will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high rewards. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

**Downloads**:

You can download the dataset file from the link given below:

**[Assignment - Dataset](https://cdn.upgrad.com/UpGrad/temp/9cefd607-bbb7-443d-bb3e-eac9fef04453/train.csv" \o "train.csv" \t "_blank)**

[file\_download](https://cdn.upgrad.com/UpGrad/temp/9cefd607-bbb7-443d-bb3e-eac9fef04453/train.csv" \o "train.csv" \t "_blank)**[Download](https://cdn.upgrad.com/UpGrad/temp/9cefd607-bbb7-443d-bb3e-eac9fef04453/train.csv" \o "train.csv" \t "_blank)**

**Data Definition**

The details of the various variables are given in the file below.

**[Data Definition](https://cdn.upgrad.com/UpGrad/temp/87f67e28-c47e-4725-ae3c-111142c7eaba/data_description.txt" \o "data_description.txt" \t "_blank)**

[file\_download](https://cdn.upgrad.com/UpGrad/temp/87f67e28-c47e-4725-ae3c-111142c7eaba/data_description.txt" \o "data_description.txt" \t "_blank)**[Download](https://cdn.upgrad.com/UpGrad/temp/87f67e28-c47e-4725-ae3c-111142c7eaba/data_description.txt" \o "data_description.txt" \t "_blank)**

The next page contains the problem statement of part-II of the assignment.

**Problem Statement - Part II**

**Assignment Part-II**

The following questions are the second part of the graded assignment. Please submit the answers in one PDF file. For writing normal text, please use MS Word (or similar software which can convert documents to PDF). For writing equations and drawing figures, you can write/draw them on a blank sheet of paper using a pen, click images and upload them in the same word document.

The final submission will be one PDF file. A sample PDF to illustrate the submission format is provided below.

**Note**: Avoid copying and pasting from *anywhere* and type the answers in your own words - your solution files will be tested using automatic plagiarism checkers and will attract a heavy penalty if plagiarism is detected.

Please limit your answers to less than 500 words per question.

**Question-1:**

Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

**Question-2:**

List at least 4 differences in detail between L1 and L2 regularization in regression.

**Question-3:**

Consider two linear models

*L1: y = 39.76x + 32.648628*

And

*L2: y = 43.2x + 19.8*

Given the fact that both the models perform equally well on the test dataset, which one would you prefer and why?

**Question-4:**

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

**Question-5:**

As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

The sample solution format is attached below.

**[Subjective Questions - Sample PDF Format](https://cdn.upgrad.com/UpGrad/temp/ebdc6ae4-ce5e-47b8-84b1-715e3da25001/Subjective+Questions+PDF+-+Sample+Format.pdf" \o "Subjective Questions PDF - Sample Format.pdf" \t "_blank)**

[file\_download](https://cdn.upgrad.com/UpGrad/temp/ebdc6ae4-ce5e-47b8-84b1-715e3da25001/Subjective+Questions+PDF+-+Sample+Format.pdf" \o "Subjective Questions PDF - Sample Format.pdf" \t "_blank)**[Download](https://cdn.upgrad.com/UpGrad/temp/ebdc6ae4-ce5e-47b8-84b1-715e3da25001/Subjective+Questions+PDF+-+Sample+Format.pdf" \o "Subjective Questions PDF - Sample Format.pdf" \t "_blank)**

The evaluation rubrics for both parts of the assignment are given on the next page.