Subjective Questions – Assignment: Advanced Regression

1. Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train accuracies, and how can this problem be solved?

Response: The model that Rahul built the training accuracy was as high as 98% but while testing the accuracy dropped to 48%, this is a classic case of overfitting. In overfitting the model performs very well on the training data but fails on the test data.

Overfitting is a phenomenon where a model becomes too specific to the data it is trained on and fails to generalise to other unseen data points in the larger domain. A model that has become too specific to a training dataset has actually 'learnt' not just the hidden patterns in the data but also the noise and the inconsistencies in the data.

Various ways to prevent overfitting are:

- **Cross-Validation** - Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model. In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set

- **Regularization** - Regularization is the process used to deliberately simplify models. Through regularization we try to strike the delicate balance between keeping the model simple yet not making it too simple to be of any use

- **Feature Reduction** - manually improve their generalizability by removing irrelevant input features

2. List at least four differences in detail between L1 and L2 regularisation in regression.

Response:

a. On Advantage of Lasso regression over Ridge Regression is that it results in model parameters, such that the lesser important features' coefficients, becoming zero. Therefore, lasso regression indirectly performs feature selection, as redundant features have a coefficient of Zero, hence they drop away.

b. Both methods cause reduction of coefficients but perform the task differently. Ridge regression, , 'sum of the squares of the coefficients', is added to the cost function along with the error term, whereas in case of lasso regression, a, 'sum of the absolute value of the coefficients is included.

L1 – Lasso Regularization

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} |\theta_i|$$

L2 – Ridge Regression

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$

c. Ridge Regression (L2) is not robust in dealing with outliers as square terms increase the error differences of the outliers and the regularization term tries to fix it by penalizing the weights, while Lasso Regression (L1) is more Robust with outliers.

d. Lasso Regression (L1) is computationally more expensive, considering the way regularization term is computed, absolute value of coefficients.

3. List at 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 data set, which one would you prefer and why?.

Response: We will prefer to use L2 which is y = 43.2x +19.8,since

- L2 is, a simpler model and such models are more generic than a complex model. We have seen that a more generic models performs better on unseen datasets.
- L2, is a simpler model and hence requires less training data. As in many cases a data scientist is provided with limited data points.
- L2, is a simple model is more robust and does not change significantly if the training data points undergo small changes and is not very sensitive to small changes.

- L2, is a simple model may make more errors in the training phase but it is bound to outperform complex models when it sees new data. A complex model is more likely to overfit.

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?

Response: To ensure the model us robust and generalisable, we use metrics which take into account both model fit and simplicity. Some such metrics are:

- Mallow's C_p
- Akaike information criterion
- Bayesian information criterion
- Adjusted R^2

All the above metrics penalise the model for being too complex and therefore help select a model that is more robust and generalizable.

Also the other method is to use Cross-validation approach. Here we keep aside some data that will not in any way influence the model building. The part of the data that is kept aside is then used as a 'proxy' for the unknown test data on which we want to estimate the performance of the mode.

Simpler and Generic models make more error in training (accuracy being low). However, these models are likely to outperform in unseen data. When a model does extremely well in training data but fails in test data, this is a case of overfitting as it has memorised the training data. While Generic and Simpler model are not too sensitive to specifics of the training data and hence higher error in the training data while perform better in test data. Simpler models have low variance and high bias.

5. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Metrics	Lasso	Ridge
R Square	84%	86%
AIC	20597.647657767953	20705.71889825625
BIC	21302.428565765502	21947.710428433747

Based on the above we have decided to select features as per Ridge

	Feature	Coeffiecient
69	Neighborhood_Somerst	4714.5
64	Neighborhood_NridgHt	5695.3
107	RoofMatl_Roll	5976.6
21	GarageCars	6396.9
106	RoofMatl_Metal	6410.6
222	BsmtExposure_Gd	6459.1
63	Neighborhood_NoRidge	6559.5
105	RoofMatl_Membran	6674.1
213	BsmtQual_Ex	9140.7
6	OverallQual	13382.4
109	RoofMatl_WdShake	13628.2
108	RoofMatl_Tar&Grv	20305.3
110	RoofMatl_WdShngl	20439.2
12	GrLivArea	32645.9
104	RoofMatl_CompShg	33177.2
0	constant	180367.4