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.

Answer: Problem occurs due to overfitting. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting, and it’s more insidious than we think.

Tackling Overfitting: Using cross validation is better and using multiple runs of cross validation is better . We want to spend the time and get the best estimate of the models accurate on unseen data.

We can increase the accuracy of our model by decreasing its complexity. In the case of decision trees for example, we can prune the tree (delete leaves) after training. This will decrease the amount of specialisation in the specific training dataset and increase generalisation on unseen data. If we are using regression for example, we can use regularisation to constrain the complexity (magnitude of the coefficients) during the training process.

Hyperparameters works well in this situation, Optimal value of lambda helps to tune rehularisation.

Question-2:

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

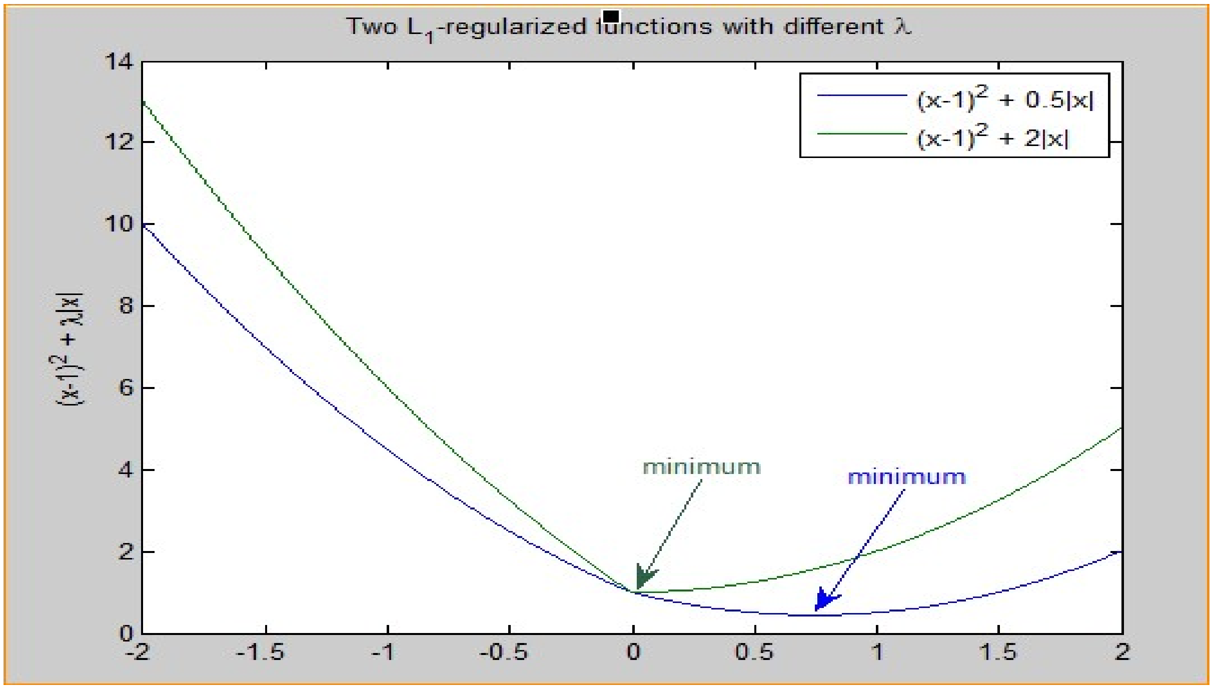
1. The main difference between L1 and L2 regularization is that L1 can yield sparse models while L2 doesn't. Sparse model is a great property to have when dealing with high-dimensional data, for at least 2 reasons.
Model compression: increasingly important due to the mobile growth

- Feature selection: it helps to know which features_are important and which features are not or redundant.

For simplicity, let's just consider the 1-dimensional case.

L2-regularized loss function $F(x)=f(x)+\lambda\|x\|22F(x)=f(x)+\lambda\|x\|22$ is smooth. This means that the optimum is the stationary point (0-derivative point). The stationary point of *F* can get very small when we increase $\lambda\lambda$, but still won't be 0 unless $f'(0)=0f'(0)=0$.

L1-regularized loss function $F(x)=f(x)+\lambda\|x\|1F(x)=f(x)+\lambda\|x\|1$ is non-smooth. It's not differentiable at 0. Optimization theory says that the optimum of a function is either the point with 0-derivative or one of the irregularities (corners, kinks, etc.). So, it's possible that the optimal point of *F* is 0 even if 0 isn't the stationary point of *f*. In fact, it would be 0 if $\lambda\lambda$ is large enough (stronger regularization effect). Below is a graphical illustration.



In multi-dimensional settings: if a feature is not important, the loss contributed by it is small and hence the (non-differentiable) regularization effect would turn it off.

2.When using L1 regularization, the weights for each parameter are assigned as a 0 or 1 (binary value). This helps perform feature selection in sparse features spaces and is good for high-dimensional data since the 0 coefficient will cause some features to not be included in the final model. L1 can also save on computational costs since the features weighted 0 can be ignored, however, model accuracy is often lost for this benefit. L1 is best used in high dimensional or sparse data sets when doing classification.

L2 regularization spreads the error among all the features. This results in weights for every feature with the possibility that some weights are really small values close to 0. L2 tends to be more accurate in almost every situation however at a higher computational cost. It is best used in non-sparse outputs, when no feature selection needs to be done, or if you need to predict a continuous output

Due to the absolute value, L1 regularization provides with a non-differentiable term, but despite of that, there are methods to minimize it. L1 regularization is also robust to outliers.

3.L2 regularization (Ridge regression) on the other hand leads to a balanced minimization of the weights. Since L2 uses squares, it emphasizes the errors, and it can be a problem when there are outliers in the data. Unlike L1, L2 has an analytical solution which makes it computationally efficient.

4.L2 regularization mainly focuses on keeping the weights as low as possible.

In contrast, L1 regularization's shape is diamond-like and the weights are lower in the corners of the diamond.

5.L2 is smooth and differentiable and L1 is sharp and non-differentiable.

In few words, L2 will aim to find small weight values whereas L1 could put all the values in a single feature.

L1 and L2 regularization methods are also combined in what is called **elastic net regularization**.

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?

Answer

We will choose the model whose R2 & adj R2 is almost nearer & higher than other model.

Here both are performing well in test data set and coefficient vale of l2 is slightly higher than l1. It may produces less p value. It may increase R2 or R2 remain same.

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?

Robust is a characteristic describing a **model's**, test's or system's ability to effectively perform while its variables or assumptions are altered, in order for a **robust** concept to operate without failure under a variety of conditions. In general, being **robust means** a system can handle variability and remain effective

When evaluating machine learning models, the validation step helps you find the best parameters for your model while also preventing it from becoming overfitted. Two of the most popular strategies to perform the validation step are the *hold-out strategy* and the *k-fold strategy*.

Briefly, a validation set is a subsection of your data that you set aside for the purpose of selecting the best parameters for your model. Central to hold-out validation, the validation set is not used in the initial training of the model and provides an independent measure of the performance of the model. Once the best parameter combination of each model has been found, the model is then retrained on the full data.

Pros of the hold-out strategy: Fully independent data; only needs to be run once so has lower computational costs.

Cons of the hold-out strategy: Performance evaluation is subject to higher variance given the smaller size of the data.

K-fold validation evaluates the data across the entire training set, but it does so by dividing the training set into K folds - or subsections - (where K is a positive integer) and then training the model K times, each time leaving a different fold out of the training data and using it instead as a validation set. At the end, the performance metric (e.g. accuracy, ROC, etc. -- choose the best one for your needs) is averaged across all K tests. Lastly, as before, once the best parameter combination has been found, the model is retrained on the full data.

Pros of the K-fold strategy: Prone to less variation because it uses the entire training set.

Cons of the K-fold strategy: Higher computational costs; the model needs to be trained K times at the validation step (plus one more at the test step).

Steps for robust models:

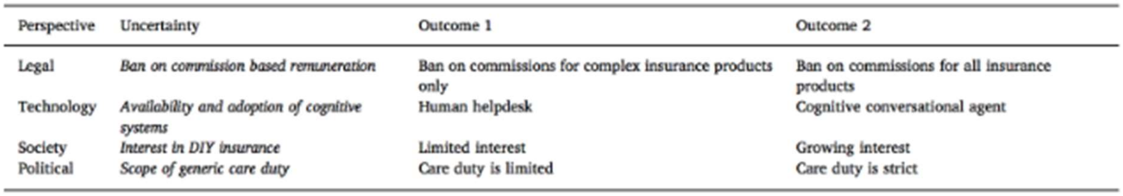
Business Model Stress-test

With the business model stress-test you confront your business model with the developments in the environment. You visualize the challenges for your business model in a 'heatmap' and discover ways to increase the robustness of your business model. This approach works with seperate component of the business model, such as value proposition, key activities or revenue model. But it also looks at the relationship between these components. The business model stresstest is done in four steps. In the first step, the business model is described and in the second step the relevant developments are collected. In step three, the actual stress test is done and step four is using the results of the stress test to improve the business models' robustness.

Step 1: <u>Describe business model</u>: Describe your business model in a structured format, such as the business model canvas or STOF business model. We will need the components of the business model to systematically evaluate it agains relevant trends and developments

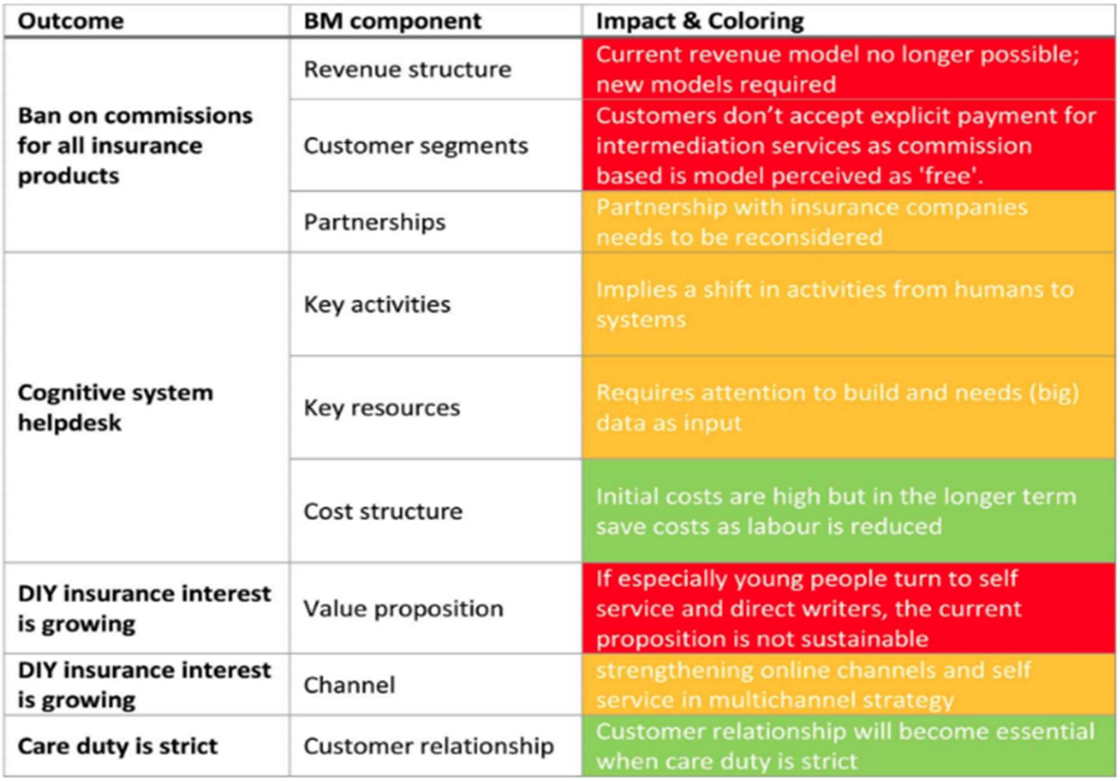
Step 2<u>: Describe the relevant trends and developments</u>: Select the most important developments that could have an impact on your business model. Keep in mind the trends of which you are certain and the developments for which the outcome is uncertain. An example of a certain trend is the aging population. Uncertain developments can be seen in macro-economics (hard/soft brexit) or legislation (how strict will privacy and data regulations become?) Reviews of trends and uncertainties can be found in trendreports business analyses, freely available on the internet. An easy way to create an overview of relevant development is a PEST analysis. Which helps cover relevant political, economical, social and technological developments. In case of uncertainties, keep in mind that there are several possible outcomes:

Summary of selected uncertainties and outcomes.



Perspective	Uncertainty	Outcome 1	Outcome 2
Legal	*Ban on commission based remuneration*	Ban on commissions for complex insurance products only	Ban on commissions for all insurance products
Technology	*Availability and adoption of cognitive systems*	Human helpdesk	Cognitive conversational agent
Society	*Interest in DIY insurance*	Limited interest	Growing interest
Political	*Scope of generic care duty*	Care duty is limited	Care duty is strict

Step 3: <u>Create a heatmap</u>: During the stresstest you confront every compontent of your business model with each of the identified developments in a 'heatmap'. To indicate the impact of a development we use colours: green for positive impact, red for negative impact with great consequences, orange to indicate the component will need to be taken into account. Make sure to write down the arguments for each colour. These are actually more important than the colour itself. For the insurance intermediary, the heatmap looks like this:



Outcome	BM component	Impact & Coloring
Ban on commissions for all insurance products	Revenue structure	🟥 Current revenue model no longer possible; new models required
	Customer segments	🟥 Customers don't accept explicit payment for intermediation services as commission based is model perceived as 'free'.
	Partnerships	🟧 Partnership with insurance companies needs to be reconsidered
Cognitive system helpdesk	Key activities	🟧 Implies a shift in activities from humans to systems
	Key resources	🟧 Requires attention to build and needs (big) data as input
	Cost structure	🟩 Initial costs are high but in the longer term save costs as labour is reduced
DIY insurance interest is growing	Value proposition	🟥 If especially young people turn to self service and direct writers, the current proposition is not sustainable
DIY insurance interest is growing	Channel	🟧 strengthening online channels and self service in multichannel strategy
Care duty is strict	Customer relationship	🟩 Customer relationship will become essential when care duty is strict

The heatmap shows the strong and weak point of your business model in relation to the external developments. The red components could cause your business model trouble. The orange components should be examined further at least. For the insurance intermediary, the uncertainty around commissions deserves extra attention for example. The question is, how an intermediary could use this development to pivot to DIY-insurance, where consumers are looking to insure themselves online instead of doing it through an intermediary.

Step 4: Analyse and improve

The last step is seeing where the weak points are and which changes could be implemented to make your business model more robust. Check the arguments that were made for choosing an orange or red colour. These can often help you formulate the specific action that should be taken to improve the business model. The insurance intermediary will want to decrease the risk of dependancy on his commissions and will need to look for new revenue models. Such as asking an hourly rate or an 'all-in' tariff for specific services or a subscriptionmodel for pro-active and long-term support. The intermediary could also experiment with the DIY-insurance methods for specific insurances and learn what the added value is for customers and where potential costs can be spared.

With the stress test we not only point out what the weak spots and risks are, but how to potentially improve the business model and make it future-proof!

Generalisable approach: The ordinary linear model is guaranteed not to fit. It will yield probabilities estimates outside [0,1][0,1]. A better approach to checking the assumptions of an ordinal regression model are:

1. First, relax the assumptions allowing for nonlinear effects using regression splines
2. Then, check the equal slopes (parallelism) assumption.

For the logistic ordinal model (proportional odds model) the equal slopes (proportional odds) assumption can be checked in several ways, including:

1. Fit a series of binary logistic models for different cutoffs of YY and plot the regression coefficients vs. cutoff and check for constancy
2. Construct partial residual plots for all cutoffs of YY. One often has to collapse infrequent YYcategories to carry this out.
3. Plot the logit of the empirical cumulative distribution of YY stratified by very important predictors and check for parallelism.

Question:What are the implications of the same for the accuracy of the model and why?

Answer: The prediction accuracy has been the long-lasting and sole standard for comparing the performance of different image classification models, including the ImageNet competition. However, recent studies have highlighted the lack of robustness in well-trained deep neura;l networks to adversarial examples. Visually imperceptible perturbations to natural images can easily be crafted and mislead the image classifiers towards misclassification. To demystify the trade-offs between robustness and accuracy, we thoroughly benchmark 18 ImageNet models using multiple robustness metrics, including the distortion, success rate and transferability of adversarial examples between 306 pairs of models. Our extensive experimental results reveal several new insights: (1) linear scaling law - the empirical ℓ_2 and ℓ_∞ distortion metrics scale linearly with the logarithm of classification error; (2) model architecture is a more critical factor to robustness than model size, and the disclosed accuracy-robustness Pareto frontier can be used as an evaluation criterion for ImageNet model designers; (3) for a similar network architecture, increasing network depth slightly improves robustness in ℓ_∞ distortion; (4) there exist models (in VGG family) that exhibit high adversarial transferability, while most adversarial examples crafted from one model can only be transferred within the same family.

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?

Answer

The lamda parameter controls the amount of regularization applied to the model. A non-negative value represents a shrinkage parameter, which multiplies P(α,β)P(α,β) in the objective. The larger lambda is, the more the coefficients are shrunk toward zero (and each other). When the value is 0, regularization is disabled, and ordinary generalized liner models are fit.

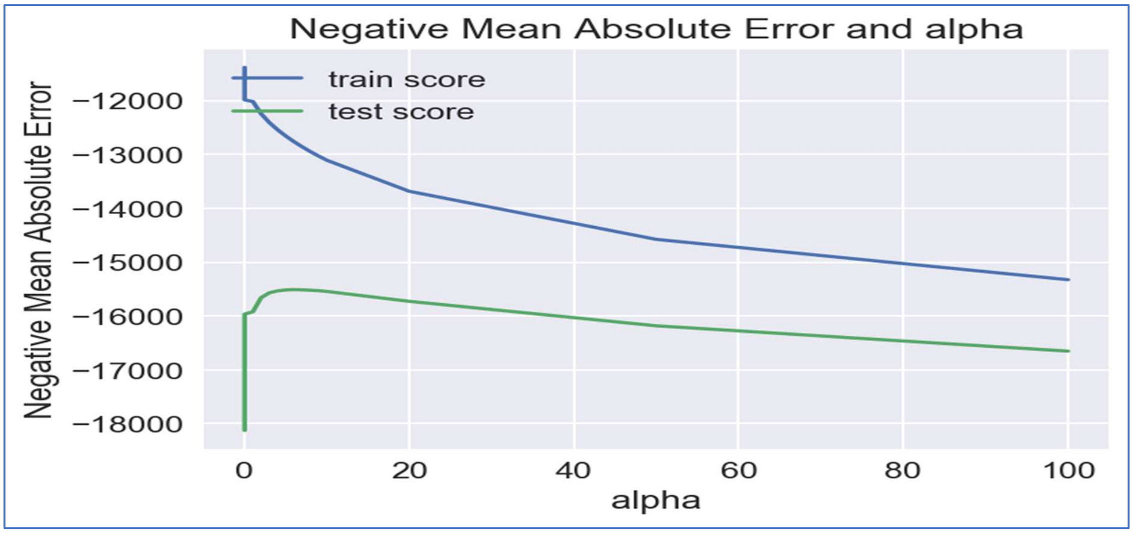
This option also works closely with the alpha parameter, which controls the distribution between the ℓ1ℓ1 (LASSO) and ℓ2ℓ2 (ridge regression) penalties. The following table describes the type of penalized model that results based on the values specified for the lambda and alpha options.

lambda value	**alpha value**	**Result**
lambda == 0	alpha = any value	No regularization. alpha is ignored.
lambda > 0	alpha == 0	Ridge Regression
lambda > 0	alpha == 1	LASSO

We taken below range for Ridge & lasso regression.

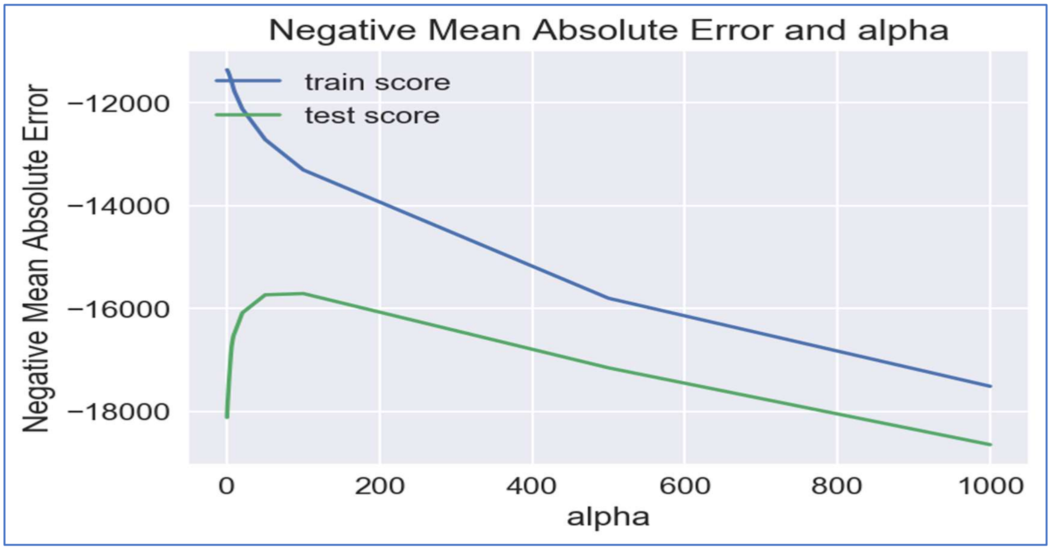
[0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000]}.

We will take the optimal value from the list for the ridge & lasso regression.



Ridge regression fig1(Alpha-15)

For ridge we take alpha - 15 & for lasso we take alpha-100 as optimal value to get better result of R2.



Lasso fig2(Alpha-100)

The graph above plots training and test error measured by Mean Squared Error. The value of lambda is varied and the associated error rates are displayed. Lower values of lambda produce the most flexible functions and these will fit the training data more accurately and therefore produce lower training error rates; however, these flexible functions may not generalize to test data. The optimal value of lambda is the value that produces the lowest test errors. So *lower the constraint (low λ) on the features, the model will resemble linear regression model.* In few words, L2 will aim to find small weight values whereas L1 could put all the values in a single feature