Regularization

1. Regularization
    a. What is it?
        i. Penalizing of coefficient because in overfitting coefficients are inflated
    b. How it works:
        i. Add penalty to model's complexity
        ii. Make model simpler – reduces num of features
        iii. Regularization works by adding a penalty term to the model's loss function, which constrains large parameter values. This constraint on parameter values helps prevent overfitting by reducing the model's complexity
        iv. If too high regularization – leads to underfitting (high bias model)
    c. Types:
        i. L1 or LASSO (Least Absolute Shrinkage and Selection Operator) –

$$\varepsilon_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 X_1 + \cdots s + \beta_n X_n))^2 + \alpha\sum_{i=0}^{n}|\beta_i|$$

            1. Can reduce coeffs to **exactly** 0 => such features are completely **discarded**
            2. Induces sparsity

$$\varepsilon_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n))^2 + \alpha\sum_{i=1}^{n}\beta_i^2$$

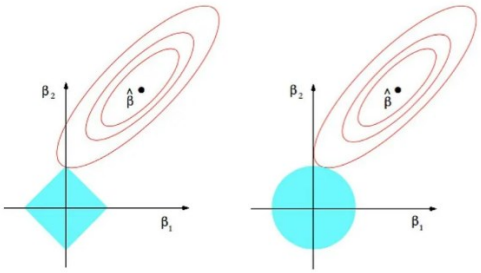
        ii. L2 or Ridge -
            1. Brings coeffs ***close*** to 0 and not exactly 0
            2. Best to use **when correlated features** are present. Because L2 regularization will evenly distribute the coefficients among those features.
            3. L2 regularization ensures that model does not become overly reliant on any single feature (unlike L1), thereby maintaining a balance in contribution of all features
        iii. Elastic net
            1. Combination of L1 and L2 regularization:

$$\text{Loss} = (1/n) * \Sigma(\text{yi} - \hat{\text{y}}\text{i})^2 + \lambda * (\text{r} * \Sigma|\text{bi}| + (1\text{-r}) * \Sigma\text{bi}^2)$$

$$Penalty_{ElasticNet} = \lambda\,(\alpha\sum|w_i| + \frac{1-\alpha}{2}\sum|w_i^2|$$

2. When to choose L1 regularization vs L2 regularization?
    a. When number of features too large – L1, because it discards unimportant features
    b. L1 induces sparsity -> faster computation, low memory usage
    c. L1 is less sensitive to outliers as compared to L2

    d. Use L1 when feature selection and model interpretability is important. Use L2 when handling model stability is important or while handling correlated features.
    e. When features have multicollinearity – L2, because it distributes the importance og correlated features more evenly
3. What is diamond shape in L1 regularization and circular shape in L2 regularization?



4. Similarity between Dropout and Random Forest?
    a. Already answered
5. Dropout
    a. How it works
        i. Randomly deactivates a fraction of neurons in a layer
    b. Why it works
        i. It simulates training multiple neural networks in parallel (similarity with RF), forcing it to be more robust and preventing it from relying too much on one neuron