Machine Learning

Assignment

1. **a**
2. **b)**
3. **c)**
4. **a),b),c)**
5. **a)**
6. **c),d)**
7. **c)**
8. **a)**
9. **a),b),c)**

10. The adjusted R-squared compensates for the addition of variables and **only increases if the new predictor enhances the model above what would be obtained by probability**. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance. Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

- The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.
- Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.
- Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called over fitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables. R-squared is expressed as a percentage between 0 and 100, with 100 signaling perfect correlation and zero no correlation at all. The figure does not indicate how well a particular group of securities is performing. It only measures how closely the returns align with those of the measured benchmark. It is also backwards-looking—it is not a predictor of future results.

Adjusted R-squared can provide a more precise view of that correlation by also taking into account how many independent variables are added to a particular model against which the stock index is measured. This is done because such additions of independent variables usually increase the reliability of that model—meaning, for investors, the correlation with the index.

11. This is a regularization technique used in feature selection using a Shrinkage method also referred to as the **penalized regression method**. Lasso is short for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator, which is used both for regularization and model selection. If a model uses the **L1 regularization** technique, then it is called lasso regression.Lasso regression follows the regularization technique to create prediction. It is given more priority over the other regression methods because it gives an accurate prediction. Lasso regression model uses shrinkage technique. In this technique, the data values are shrunk towards a central point similar to the concept of mean. The lasso regression algorithm suggests a simple, sparse models (i.e. models with fewer parameters), which is well-suited for models or data showing high levels of multicollinearity or when we would like to automate certain parts of model selection, like variable selection or parameter elimination using feature engineering.

Lasso Regression Implementation in Python using sklearn

```
from sklearn.linear_model import Lasso

lassoReg = Lasso(alpha=0.3, normalize=True)

lassoReg.fit(x_train,y_train)

pred = lassoReg.predict(x_cv)

# calculating mse

mse = np.mean((pred_cv - y_cv)**2)
```

Lasso Regression algorithm utilises L1 regularization technique It is taken into consideration when there are more number of features because it automatically performs feature selection.

RIDGE REGRESSION

Ridge Regression is another type of regression algorithm in data science and is usually considered when there is a high correlation between the independent variables or model parameters. As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value. Therefore, we create a bias matrix in the equation of Ridge Regression algorithm. It is a useful regression method in which the model is less susceptible to overfitting and hence the model works well even if the dataset is very small.

Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square. Ridge regression is also referred to as **L2 Regularization**.

Ridge Regression Implementation in Python using sklearn

```
from sklearn.linear_model import Ridge


## training the model

ridgeReg = Ridge(alpha=0.05, normalize=True)

ridgeReg.fit(x_train,y_train)

pred = ridgeReg.predict(x_cv)

calculating mse

mse = np.mean((pred_cv - y_cv)**2)

print(mse)

## calculating score

score = ridgeReg.score(x_cv,y_cv)

print(score)
```

12. A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

KEY TAKEAWAYS

- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.
- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.
- The formula for VIF is:

- Where R_i^2 represents the unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones.

13. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. As we know most of the supervised and unsupervised learning methods make decisions according to the data sets applied to them and often the algorithms calculate the distance between the data points to make better inferences out of the data.

14. Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at MSE, MAE, R-squared, Adjusted R-squared, and RMSE.

Mean Squared Error (MSE)

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Mean Absolute Error (MAE)

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

R-squared or Coefficient of Determination

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.To understand what R-square really represents let us consider the following case where we measure the error of the model with and without the knowledge of the independent variables.

Root Mean Squared Error (RMSE)

This is the square root of the average of the squared difference of the predicted and actual value.R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure).Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

15. Confusion matrix calculation:

Actual/predicted	True	False
True	**1000**	**50**
False	**250**	**1200**

Sensitivity formulae=True positive\ (True positive+False negative)

True positive rate (TRP) = 1000/ (1000+250)

=0.800

Specificity formulae=TN/ (FP+TN)

SPC=1200/ (50+1200)

=0.9600

Precision formulae=TP/ (TP+FP)

PPV=1000/ (1000+50)

=0.8276

Accuracy formulae= (TP+TN)/(P+N)

ACC= (1000+1200)/(1200+1200)

=0.8800