**Linear Regression Report**

Student’s Name

Institution

Course Name & Number

Instructor’s Name

Date

**Description of the problem and dataset**

The dataset contains data on a social network advertising campaign company. It contains 400 observations on the age, estimated salary of users and whether they made a purchase or not. The objective of conducting an analysis on the data is to understand the factors influencing purchase decisions of users.

**Selected Method (Logistic` Regression)**

For my analysis, I selected the logistic regression algorithm. Logistic regression is a statistical method used for binary classification tasks. It is a predictive analysis used to describe data and explain the relationship between one dependent binary variable (e.g., purchased column) and one or more ordinal independent variables.

**Key Concepts of Logistic Regression**

1. **Binary outcome –** Logistic regression is suitable for modelling binary outcomes, where the dependent variable can take on one of two values, normally 0 or 1.
2. **Outliers –** There should be no outliers in the data, which is possibly assessed by converting the continuous predictors to standardized scores.
3. **Multicollinearity –** There should be no high correlations among the predictors. This can be assessed by a correlation matrix among the predictors.
4. **Overfitting –** Adding independent variables in a logistic regression model will always increase the amount of variance (R2). Adding more and more variables to the model can result in overfitting which reduces the accuracy of the model beyond the data on which the model is fit.

**Data Preparation**

The provided dataset was prepared as follows:

1. Data Cleaning – The data was checked for missing values which were to be replaced with zero if found. However, the dataset had no missing values and was clean.
2. Checking for outliers – The data was checked for outliers, which are observations that significantly deviate from the rest of the data distribution. These were mainly applicable in the dependent variable (Purchased) as it was supposed to hold either 0 or 1 values. The dataset however had no outliers.

**Training Approach and parameter selections**

The dataset was split into training and testing sets using a test size of 25% and a training size of 75%. The random state used was 42, ensuring the model’s performance could be evaluated on the unseen data. To maintain uniformity and scale, the features in the dataset were standardized using the “Standard Scaler” library from scikit-learn. This was essential to ensure that each feature contributed proportionately to the model’s predictions and prevented some features from dominating others during training.

The logistic regression model was then trained using the “fit ()” method of the logistic regression class of scikit – learn. The training process involved learning parameters that best fit the training data, allowing the model to predict the probability of a user making a purchase based on their age and estimated salary.

**Parameter Selections**

During the training process, default parameters were used for the logistic regression model. These include:

* Regularization strength (C): This parameter determines the amount of regularization applied to the model and was set to 1.0 by default.
* Solver: This parameter represents the algorithm used for optimization and is set to “lbfgs” by default.
* Max iterations: Refers to the maximum number of iterations taken for the solver to converge and is 100 by default.

Manual parameter tuning was not done due to the satisfactory performance of the model with the default parameters.

**Model evaluation results and metrics**

The logistic regression model displayed impressive performance in predicting the purchases made by users depending on their age and estimated salary. The following metrics demonstrate the performance:

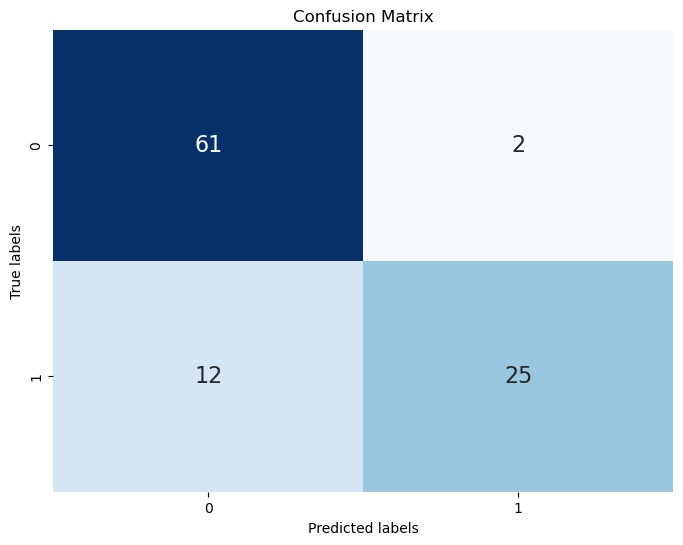
* Accuracy – The model had an accuracy of 86%, meaning it correctly classified 86% of the instances in the test set.
* Precision – The precision score was 92.6% which means it accurately identified almost 93% of the purchases provided as positive, out of the total purchases predicted as positive.
* Recall – The recall score for the model was 67.57% indicating that the model correctly identified 67.57% of the actual positive purchases in the dataset.
* F1 Score – The F1 score had a value of 78.13%. This metric provides a balance between precision and recall ensuring that both false positives and false negatives are considered in the model’s performance assessment.
* Confusion Matrix – Provides a detailed breakdown of the model’s predictions as compared to the actual values. The confusion matrix for the model was:
  + True positive (TP): 25 instances were correctly classified as purchases.
  + False positive (FP): 2 instances were incorrectly classified as purchases
  + True Negative (TN): 61 instances were correctly classified as non-purchases.
  + False Negative (FN): 12 instances were incorrectly classified as non-purchases.

**Findings of study and model results**

The logistic regression model exhibited promising performance in classifying users into purchasers and non-purchasers. The model had an accuracy of 86% and a precision score of 92.6%, indicating that the model was highly successful in predicting sales. However, its recall score of 67.57% indicates it had a moderate level of detecting actual positive instances from the dataset.

The findings suggest that factors such as age and estimated salary are significant in predicting user purchases on social networks. The model could further be improved by optimizing model parameters manually and techniques such as cross validation.

**Visualisations of algorithm performance**



**Conclusion**

The model illustrated impressive performance in predicting whether a user would make a purchase based on their age and estimated salary, achieving an accuracy of 86% and a precision of 92.6%. However, the model’s recall score was 67.57%, indicating areas of improvement needed in capturing potential purchasers. The confusion matrix revealed false negative instances, suggesting opportunities for further optimization.