**Assignment 1: Customer Subscription Prediction with Machine Learning**

**Introduction**

1. **Can you build a model to predict whether a customer will subscribe to a term deposit?**
2. Yes, we can build a machine learning model to predict whether a customer will subscribe to a term deposit. We can use Logistic Regression, Decision Tree, Random Forest Algorithms.

**Q) What are the most important factors that influence a customer's decision to subscribe?**

A) The features like job, marital status, education, housing loan, other loans, previous campaign data are important factors that influence a customer's decision to subscribe

**Q) What are the limitations of your model?**

**A) The following are the limitations of logistic regression model:**

* There is a chance of over-fitting if dealing with high-dimensional data
* It may not always capture the patterns and there are much better, efficient algorithms.
* This model is sensitive to outliers, so making sure there is absence of outliers.

**Exploratory Data Analysis (EDA)**

**Q) What is the distribution of the customer ages?**

A) The ages vary from a little less than 20 to above 90 years old. The mean being 40.936 years. The age with the most counts is 32 years and the least being 94 years.

**Q) What is the relationship between customer age and subscription?**

A) From the correlation matrix, the correlation is 0.025155. which is positive correlation yet weak. Hence, we conclude that age doesn’t really play a major relevance compared to other features.

**Q) Are there any other factors that are correlated with subscription?**

A) Yes, there are other features like job, marital status, education, housing loan, and previous campaign data which correlate with subscription.

**Conclusion**

The customers are more likely to subscribe to the term deposit

* if they have more average yearly balance.
* The longer they speak to telemarketers
* If they are contacted in march, October and September ( Based on correlation matrix values)

**Limitations:**

1. The unknown values present in the data set.
2. High chance of overfitting.
3. Smaller changes in data set might lead to huge differences in random forest and decision trees.

**Questions:**

1. **What is the distribution of the customer ages?**

A) The ages vary from 18 to 95. The mean of the ages being 40.936.

**2. What is the relationship between customer age and subscription?**

A) From the correlation matrix, the correlation is 0.025155, which is positive correlation yet weak.

**3. Are there any other factors that are correlated with subscription?**

A) Yes, there are other features like job, marital status, education, housing loan, and previous campaign data which correlate with subscription.

**4. What is the accuracy of the logistic regression model?**

A) The accuracy of the logistic regression model is 88.88

**5. What are the most important features for the logistic regression model?**

A) The most important features for the logistic regression model is job, marital, education, housing, loan, contact, month and poutcome.

**6. What is the precision of the logistic regression model?**

A) The precision for the class 0 is 0.90 and for class 1 is 0.61 for logistic regression model.

**7. What is the recall of the logistic regression model?**

A) The recall for the class 0 is 0.98 and for class 1 is 0.22 for logistic regression model.

**8. What is the f1-score of the logistic regression model?**

A) The f-1 score for the class 0 is 0.94 and for class 1 is 0.32 for logistic regression model.

**9. How can you improve the performance of the logistic regression model?**

**A)** One can improve the performance of the logistic regression model by experimenting with different combination of features and better quality of data set (no missing values and preferably less number or no unknown values).

**10. What are the limitations of the logistic regression model**

**A)** The following are the limitations of logistic regression model**:**

* It assumes that predictors are independent of each other, and violations of this assumption can lead to biased coefficient estimates.
* Logistic regression can be influenced by outliers
* logistic regression may become computationally expensive and prone to overfitting.
* The model does not handle missing data well, necessitating data imputation or other techniques.
* The model's linearity may not effectively capture intricate interactions and complex patterns in the data.