Project Title

(You are expected to give a descriptive title)

Contents

1 Introduction

This section is expected to address the following information:

- Some background information about the project: objective and motivation.
- Introduction to the working data set: sample size information about data collection, variables description (names, description/definition/type), etc.
- Clear statements of questions to be addressed (make sure both linear and logistic regression modeling will be used to address the questions).

2 Methodology

Since this is a comprehensive analysis report, you are expected to use a standalone section to outline the methods and models that will be used to address the questions. Note that you are expected to write several narrative paragraphs to describe each of the potential models and their assumptions.

You can create subsections to describe individual models/algorithms.

3 EDA and Feature Engineering

Feature engineering based on EDA. Other algorithm-based engineering methods are optional. If you have prior experience with any model-based algorithms, please feel free to use them to make powerful variables. Since this is a project focusing on statistical analysis, you need to consider the interpretation of regression models (i.e., coefficients). When using transformations, you need to think about the interpretation of the model.

3.1 Handling Missing Value

If your data set has missing values, you need to use appropriate imputation methods to handle missing values.

3.2 Single Variable Distribution

Based on observed patterns from the visualizations, you may take appropriate actions such as regrouping, binning, discretization, etc. to make more powerful variables for subsequent models and algorithms.

3.3 Assessing Pairwised Relationship

Through observing the potential correlations decide whether to drop highly correlated variable(s).

4 Linear Regression Modeling

You are expected to follow the general model-building process to search for the final model to address related questions. Please include any visual representations whenever possible (see how visualizations were included in lecture notes).

4.1 Create Candidate Models

See the section of the project guidelines.

4.2 Use Cross-validation for Model Selection

Use cross-validation and MSE as a predictive performance measure to select the best model.

4.3 Results and Conclusions

Report results based on the final model and conclude the statement of questions.

5 Logistic Regression

Use the same model-building process as you did in the linear regression model to search for the best logistic regression model for prediction.

5.1 Create Candidate Models

See the section in the project guidelines.

5.2 Model Selection

Using ROC curve and AUC to compare candidate models. Both ROC curves and AUC need to be included in the report

5.3 Cut-off Probability Search

Identify the optimal cut-off probability that yields the best prediction accuracy,

5.4 Results and Conclusion

Report the results and conclusion of the logistic regression.

6 Summary and Discussion

Summary of the project and discuss the strengths and weakness of the methods and potential improvements.

7 Reference and Appendix

List all references cited in the project. Add any appendices you may have to support any arguments in the report.