End-of-Quarter Report

This quarter, I focused on mastering key concepts and techniques in statistical learning, with a particular emphasis on linear regression, classification methods, and resampling techniques. Below is a review of the topics I learned and the practical project I undertook to synthesize these concepts.

Topics Covered

1. Linear Regression:

Simple Linear Regression: Learned to estimate coefficients and assess their accuracy, ensuring robust interpretations of relationships between predictors and response variables.

Multiple Linear Regression: Explored techniques to handle multiple predictors, including interaction terms, qualitative predictors, and non-linear transformations.

Model Evaluation: Analyzed model accuracy using metrics like Mean Squared Error (MSE) and addressed potential issues like multicollinearity.

Lab Exercises: Practiced implementing these techniques with real-world datasets, focusing on multivariate goodness-of-fit and predictor interactions.

2. Classification

Logistic Regression: Studied binary classification through the logistic model and its extensions to multinomial logistic regression.

Discriminant Analysis: Explored Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) for generative modeling of classification problems.

K-Nearest Neighbors (KNN): Learned about flexibility trade-offs, balancing high bias and variance for optimal model performance.

Comparison of Methods: Conducted empirical and analytical comparisons of various classification techniques.

3. Resampling Methods

Cross-Validation: Applied validation set approaches, leave-one-out cross-validation (LOOCV), and k-fold cross-validation to evaluate and optimize models.

Bootstrap Methods: Practiced estimating variability and confidence intervals for predictions, enabling better model reliability.

Special Project: Titanic Survival Analysis

To consolidate my learning, I applied these methods to analyze the Titanic dataset. This project aimed to predict passenger survival using logistic regression, feature engineering, and cross-validation. Highlights include:

- **Model Training: Implemented logistic regression with polynomial features of degree 3, using variables like Pclass, Sex, Age, and Fare.**
- **Cross-Validation: Used 5-fold cross-validation to identify the optimal model complexity, achieving an accuracy of 81.56% and an ROC-AUC score of 87.31%. These metrics demonstrated a strong balance between model complexity and performance.**

Key Takeaways

- **The interplay between bias and variance emerged as a critical theme, underscoring the importance of balancing model complexity for better generalization.**
- **Resampling techniques like cross-validation and bootstrapping proved invaluable for validating model reliability and avoiding overfitting.**
- **Feature engineering and thoughtful variable selection can significantly enhance predictive performance, emphasizing the importance of domain knowledge in data science.**

Moving forward, I plan to explore advanced statistical and machine learning techniques, such as ensemble methods and deep learning, to address more complex datasets. And focus on developing interpretable models that ensure fairness and accountability especially in applications related to public health and policy. Most importantly, I am seeking collaborative research opportunities to expand my practical experience and contribute to impactful projects.

This Directed Reading Project provided an excellent foundation for understanding statistical learning methods and their practical applications. The knowledge and skills gained have reinforced my commitment to leveraging data science for real-world problem-solving.