Hugh_Reflection

2023-11-08

Reflection for Assessment 1, Week 7

Hugh Pullin, Nov 2023

Our group was assigned an open binary classification problem, with a strong emphasis on selecting an appropriate performance metric for model evaluation.

Our objectives were:

1. To each submit a model to be ran on left-out test data
2. Agree on and compare the models with a chosen performance metric

Learning from Assessment 0, we were clear about the type of data we needed: numerous features, a large number of data points, and a definitive binary classification problem. Initially, I proposed several datasets, eager to rectify the issues of the previous assessment. However, due to various learned reasons - such as inadequate entries, a weak binary classification problem, and highly correlated features — we didn't have a good dataset until we found 'smoking_drinking_data.csv.' With 22 features, both ternary and binary classification, and nearly 1 million rows, this dataset seemed ideal for our task.

Establishing a robust foundation for the project was a significant priority for me. Recognizing the importance of the chosen performance metric, I aimed to contextualize the classification problem to better select the performance metric. We framed our scenario as health insurance brokers, evaluating the implications of each performance metric within this context. Consequently, we identified sensitivity as our preferred performance metric. As brokers, correctly identifying drinkers and assigning higher premiums for associated health risks was critical. A false negative (assigning a lower premium to a drinker) could be rectified via reimbursement of the policy holder, while a false positive (failing to assign a higher premium to a drinker) would incur potential financial loss if we were to cover a health issue we hadn't collected sufficient funds to pay for. Therefore, we determined the true positive rate to be the key performance indicator for our models. I was glad to pioneer this approach allowing us to establish a concrete framework for comparing all the models against this performance metric.

Subsequent to data description and exploratory data analysis (EDA), we used plots and descriptive functions to comprehend the data. While I previously deemed EDA as important but not critical, I realized its significance after multiple iterations upon discovering previously unexplored features. With my expert knowledge, I intend to conduct a more rigorous EDA in the next assessment.

Working with 23 features and understanding the significance of feature selection from the literature I read to support my work[1], I recognised the necessity of selecting covariates. Despite attempting feature reduction through correlation matrix plotting and defining a correlation threshold function, I couldn't produce effective results. I avoided working backward from the results to create a smaller yet effective feature subset as I deemed this to be malpractice. Instead, I utilized PCA to create a second dataset for testing. Contrary to what was observed in literature and example code, the tree trained on PCA-transformed data performed worse, indicating an insufficient capture of variance. Despite this contrary outcome, it was important to highlight this discrepancy.

My choice of the decision tree as the model stemmed from my initial understanding of the concept in lectures, with the interpretability via dendrogram appearing suitable for the assessment. Later, I could discuss the efficiency and scalability of decision trees in a real-world context. Ultimately, the decision tree achieved the highest true positive rate. The decision tree is a robust model that uses Gini impurity as a splitting criterion to minimize the probability of incorrectly classifying an element. The recursive process of splitting data continues until meeting the stopping criterion.

If the intention were to "win," I would maintain most decisions - given mine was the model to "win". Despite attempting to optimize the decision tree factors, which produced consistent results, I'd focus more on data processing (feature engineering) and its influence on further improving sensitivity.

On the whole, I am content with my progress since the previous assessment, especially regarding my overall contributions to coding and project direction. We navigated the project reasonably well; however, I faced challenges delving deeper into specific issues encountered with the data, such as the accuracy reduction when using PCA-transformed data or optimizing the decision tree at its lowest complexity. Further exploration is needed to address these concerns, although we provided explanations for what we believed were the underlying causes.

To enhance, a greater emphasis on understanding the data and the relationships between features would have been crucial to better equip us in resolving issues in the future.

[1] Kumar (2020) "Decision Trees for Binary Classification"