Tania Lopes

Student No. 22202398

INDIVIDUAL REFLECTION REPORT

1. PERSONAL CONTRIBUTION

As part of a two-person team for this project, my colleague and I divided responsibilities to leverage our strengths and ensure the project's smooth completion. In the initial stages of the project, I took the lead on data preparation. This involved running the two datasets we had, mine and my teammate's, through a data preparation notebook. I then embarked on a detailed analysis to identify correlated features by plotting categorical and continuous features against our target features. The objective of this analysis was to decide which features to retain and which to drop, a decision reached with input and discussion from my teammate. Part of this discussion involved considering the potential for multicollinearity and identifying the features that could cause this in our model.

The next phase of our project involved modeling, which my teammate and I shared. She handled the linear and logistic regression models, while I took responsibility for the random forest model. Our approach was greatly facilitated by the reuse of code from practical exercises and solution sets provided by our professor. My software development contribution primarily involved encapsulating the existing code into different functions for easier reuse. These functions were particularly useful for printing evaluation metrics and producing models quickly when testing optimizations.

While the majority of the COVID-specific domain research had been conducted during a prior assignment, I did undertake substantial domain research into the models. My focus was predominantly on the random forest model, given my assigned responsibility, and later, I delved into logistic regression when my teammate encountered issues that led to her model consistently predicting the majority class.

In the final stages of the project, I spearheaded the optimization process and its analysis. The objective here was to fine-tune our models to improve their predictive accuracy. I thoroughly explored different optimization techniques, testing each one and analyzing their impacts on our model's performance.

Finally, I dedicated significant effort to structuring our report and ensuring it met all the requirements laid out in the assignment brief. My goal was to make our report comprehensive, coherent, and easy to follow, reflecting the care and effort put into the modeling and interpretation of the results.

2. SKILLS AND LESSONS LEARNED

This project served as a comprehensive learning experience, equipping me with new skills and valuable insights about predictive modeling.

A pivotal skill I honed during the project was understanding and implementing decision trees, the foundational algorithm behind the Random Forest model. Grasping the concept of decision trees was instrumental in

realizing how they partition data to create pure subsets. It also highlighted the potential for overfitting due to excessive complexity, a pitfall that can be mitigated by setting maximum depth limits.

The importance of feature selection came to the fore during the project. Proper selection of features is vital for enhancing model performance. However, I learned that it's a balancing act. Too many features can lead to overfitting, while too few might result in a model that underperforms due to insufficient information. This understanding will guide me to be more judicious about feature selection in future projects.

Data formatting was another area where I gained significant knowledge. I realized how essential it is to prepare data in a way that is compatible with the model. Through normalizing the data, the logistic regression model was able to yield meaningful results. The technique of dummy variable encoding and the necessity of dropping one dummy variable to avoid multicollinearity were also crucial lessons.

One of the critical issues I faced was class imbalance in the target variable. The model was better at predicting survival instances due to their higher prevalence in the dataset. Understanding and addressing this imbalance through oversampling and undersampling techniques was a key learning point. I also came to appreciate the concept of diminishing returns in machine learning. There is a point beyond which adding more features can be counterproductive, leading to unnecessary complexity and potential overfitting. This insight underscored the need for strategic feature inclusion, a lesson I'll carry into future projects.

Reflecting on the project, certain aspects went particularly well. For instance, the initial data preparation and exploratory analysis enabled us to gain a deep understanding of the dataset. This understanding, in turn, informed our choices during the modeling process. Moreover, I also thought the iterative nature of the optimizations went really well. It allowed us to see the effects of all of our decisions, and refine our models according to what was most effective.

There were areas that could have gone better. The next time I take on a project of this nature, I would like to dig deeper into the effects of categorical and continuous features on the different models, particularly a feature like state_fips_code. Despite being treated as a continuous feature, it ended up being one of the more influential features in the Random Forest model, where it was synthetically partitioning data based on the numeric value of the code.

In future projects, I would place more emphasis on more stringent feature selection at the outset, taking a more strategic approach to including only those features that provide significant predictive power. I would also pay more attention to class imbalance, exploring additional techniques such as synthetic minority over-sampling (SMOTE) to manage this issue more effectively.

This project was a stepping stone into the world of predictive modeling, imparting vital skills and lessons that will be instrumental in navigating future data analytics projects.

3. Additional Feedback

A couple of additional comments are on my initial impression of the dataset, the benefits of working with a partner, and the simultaneous feeling that while we learned a lot in doing this assignment, there is a lot more to know.

The dataset, at first glance, seemed subpar and rife with missing and inconsistent entries. Yet, it was surprising to witness the performance of our predictive models on this seemingly inadequate data. This served as a practical testament to the prowess of machine learning, even when handling imperfect data.

The project offered a hands-on introduction to machine learning, exposing the complex layers of feature selection and model optimization. Despite the challenges this complexity presented, it also fostered an intellectual curiosity and a desire to delve deeper into machine learning's intricacies.

Working collaboratively added a new dimension to the project. The process of negotiating decisions about feature importance and model optimization strategies, while demanding, was also enlightening. It was interesting to see how two individuals could approach the same problem differently, leading to diverse ideas and perspectives.

Overall, the assignment was challenging, but I feel that it cemented a lot of the things taught in the lectures and in the practicals.