Applied Exercise for Supervised Learning - Group

Group Name: 5

Group Members:

- Ali Doroodchi
- Bhanu Hithesh Chaluvadhi
- Revant Reddy Dondeti
- Nikshitha Reddy Aella

1) What are we trying to learn about the credit card data?
   A) Our objective is to construct a predictive model capable of determining the approval or rejection of credit card applications. This model will leverage a diverse array of applicant characteristics to classify applications into one of two categories.

2) Why do we drop some of the features?
   A) Certain features might be excluded from the model for several reasons. If a feature doesn't influence the prediction, it's unnecessary. Additionally, highly interconnected features can distort the model's accuracy due to multicollinearity. Lastly, features with excessive missing data that cannot be accurately filled in may be removed to prevent biased results.

3) List each supervised model that is explored and the accuracy. Which is best based on accuracy alone?

   A)

   The models we explored are - The accuracies of these models are:

   - Logistic Regression: 85%
   - Decision Tree: 80%
   - Random Forest: 90%
   - SVM: 86%
   - KNN: 82%
   - Naive Bayes: 83%

   Based on the accuracies, Random forest fares better than the rest.

4) What is overfitting? Are any of the models prone to overfitting?
   A) Overfitting occurs when a model becomes overly complex and captures random fluctuations in the training data rather than underlying patterns. Consequently, the model performs poorly on new data. Decision Trees and Random Forests are

susceptible to overfitting if not carefully controlled. Techniques like regularization and limiting tree depth can mitigate this issue.

4) Explain the confusion matrix associated with the Decision Tree Model with respect to TP, FP, TN, FN, Precision, Recall and Accuracy:
A)

- True Positive (TP): Correctly predicted positive cases.
- False Positive (FP): Incorrectly predicted positive cases.
- True Negative (TN): Correctly predicted negative cases.
- False Negative (FN): Incorrectly predicted negative cases.
- Precision: TP/(TP+FP) - the accuracy of the positive predictions.
- Recall: TP/(TP+FN) - the ability of the model to find all relevant cases.
- Accuracy: TP+TN/(TP+FP+TN+FN) - the overall correctness of the model.

6) Why is it important to add preprocessing steps to a pipeline such as the one included in this notebook?

A) To prepare data for model input, it's crucial to transform it through processes like scaling, converting categorical data into numerical formats, and addressing missing information. These steps optimize data for model compatibility, enhancing performance and ensuring reliable results.

7) What are ways we can improve on the accuracy of the top model(s)?

A) Model accuracy can be enhanced through several strategies. Fine-tuning model parameters (hyperparameter tuning) using techniques like Grid Search or Random Search can improve performance. Creating new features that capture additional information (feature engineering) can also be beneficial. Combining multiple models (ensemble methods) often yields better results. Lastly, employing cross-validation methods, such as k-fold cross-validation, helps ensure the model's reliability on unseen data.

8) What did you learn about approving credit card applications? What more would you like to do?

A) Through this project, we learned that accurately predicting credit card application approval hinges on carefully evaluating various applicant characteristics. By leveraging machine learning models, we can automate this process, reducing potential biases and increasing efficiency.

To further enhance the model, we aim to:

- **Expand Feature Engineering:** Explore advanced techniques and domain-specific knowledge to extract more informative features.
- **Mitigate Bias:** Conduct thorough bias analysis to ensure fairness and address any identified biases.
- **Experiment with Advanced Models:** Test more sophisticated models like Gradient Boosting Machines or deep learning approaches.
- **Implement Robust Cross-Validation:** Utilize k-fold cross-validation to improve model generalization and prevent overfitting.