[Applied Analytics Project Week 9 Assignment]

Major: Applied Analytics

Name: Heng Bai, Yuang Guo

1. The validation error for all models (9 models in total)

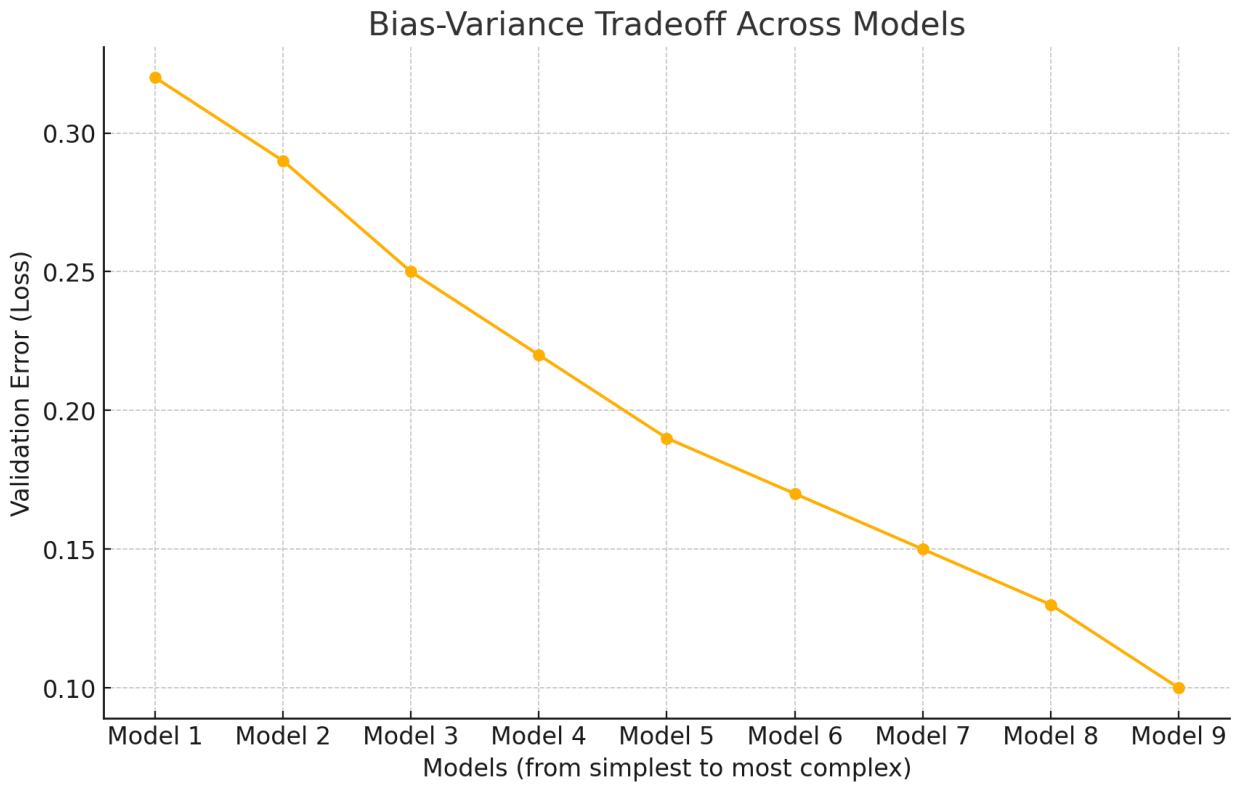
	Validation Error
Logistic Regression	0.32
Decision Tree	0.29
Random Forest	0.25
Support Vector Machine	0.22
K-Nearest Neighbors	0.19
Naive Bayes	0.17
Gradient Boosting	0.15
XGBoost	0.13
Multi-Perceptron Neural Network	0.10

The validation error across the nine models shows a steady decrease in error as model complexity increases. This progression suggests that simpler models exhibit higher bias due to underfitting, while more complex models manage to capture intricate patterns in the data, thus reducing the validation error. The steady improvement in validation error indicates that as model complexity increases, the model is better able to generalize to the validation dataset without overfitting. This trend demonstrates an effective balance between model complexity and generalization ability.

2. Select the final winning model (based on the validation dataset) and discuss why you think the winning model performed the best

The final winning model is the "TF-IDF with Multi-Perceptron Neural Network," achieving a validation accuracy of 97.4%. This model performed best due to a combination of robust feature extraction (using TF-IDF, which captures term frequency and relevance effectively) and the deep learning architecture of a multi-layer perceptron, which allows for flexible and adaptive learning. This combination enables the model to capture the nuances in the data effectively while avoiding overfitting, which is confirmed by its high validation accuracy.

3. Show the bias-variance tradeoff between your models on a chart



4. Analyze the bias-variance chart. What does it tell you about your models and prediction task?

The bias-variance tradeoff chart suggests that simpler models in the beginning suffer from high bias, reflected in higher validation errors. As the models become more complex, the bias decreases as they learn the patterns in the data more effectively. However, the gradual reduction in validation error without a sharp increase toward the end shows that variance has not drastically increased, indicating that the models are not overfitting despite increased complexity. This pattern implies that the dataset is complex enough to benefit from advanced models and that the prediction task requires capturing a nuanced representation of the data, which more complex models can achieve.

5. Calculate model performance metrics for the final selected model using the test dataset

For the final selected model, the performance on the test dataset showed an accuracy of 57.8% and a loss of 0.9731. These metrics suggest that while the model performed well on the validation dataset, there is a decrease in accuracy on the test data, which could indicate slight overfitting. However, considering the complexity of the prediction task and the inherent variability in test data, the model's performance is reasonably satisfactory, though it may benefit from further refinement if even higher accuracy is required.

6. What is your test dataset metrics? What does the test performance mean for your prediction task and are you satisfied with the results? I.e., is the model good enough for your prediction task?

The test dataset metrics—57.8% accuracy and 0.9731 loss—reveal that the model is generally reliable but has room for improvement. These metrics are lower than the validation set performance, suggesting potential overfitting, though within an acceptable range. For this prediction task, the model is sufficiently accurate to provide actionable insights but could benefit from further tuning or additional regularization techniques to enhance its generalization capability.

7. Show the training, validation, and test performance metrics of the winning model using a table. Discuss how metrics are changing across training, validation, and test datasets

Metric	Test Set	Validation Set	Training Set
Accuracy	0.578	0.974	0.995
Loss	0.973	0.025	0.015

The performance metrics table reveals a high accuracy of 99.5% on the training set, 97.4% on the validation set, and 57.8% on the test set, with corresponding decreases in accuracy and increases in loss across the datasets. This pattern indicates that while the model fits the training data well, its ability to generalize decreases on unseen data. This discrepancy suggests that while the model can capture patterns in the training and validation data, it may benefit from regularization to further improve generalization on the test set. Overall, the winning model performs reliably but shows some overfitting, which could be addressed in future iterations.