Impact of Feature Selection on Bias and Variance

Feature selection plays a crucial role in balancing bias and variance in a model:

- **Bias:** Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. A model with high bias makes strong assumptions about the underlying relationship between features and the target variable.
- **Variance:** Variance refers to the model's sensitivity to small fluctuations in the training data. A model with high variance is highly influenced by the training data and may not generalize well to new data.

Here's how feature selection affects bias and variance:

- **Removing irrelevant features:** Removing features that are not related to the target variable can help reduce variance by simplifying the model and making it less sensitive to noise in the data. This can potentially increase bias if important information is lost.
- **Removing redundant features:** Removing features that are highly correlated with each other can also help reduce variance and improve model interpretability.
- **Adding relevant features:** Including features that have a strong relationship with the target variable can help reduce bias by allowing the model to capture more complex patterns. This can potentially increase variance if the added features introduce noise or make the model too complex.

In our experimentation, we trained a model with a subset of features ('Pclass' and 'Fare'). The R^2 score for this subset model ({{r2_subset}}) was lower than the R^2 score for the model trained with all features ({{r2}}). This suggests that removing features in this case increased the bias of the model, as it was less able to explain the variance in the target variable. The simpler model with fewer features likely has lower variance, but at the cost of increased bias. The goal of feature selection is to find the optimal balance between bias and variance to achieve the best possible model performance on unseen data.

Linear Regression Model Summary Report

This report summarizes the key findings from the linear regression model trained on the Titanic dataset.

Key Results and Performance Metrics

- **Mean Squared Error (MSE):** {{mse}}
- **Root Mean Squared Error (RMSE):** {{rmse}}
- **R-squared (R^2) on Training Set:** {{r2_train}}
- **R-squared (R^2) on Testing Set:** {{r2}}
- **R-squared (R^2) with Pclass and Fare features:** {{r2_subset}}

The MSE and RMSE values indicate the average magnitude of the errors in the model's predictions. The R^2 scores represent the proportion of the variance in the target variable ('Survived') that is predictable from the features.

Overfitting/Underfitting Observation

Comparing the R^2 on the training set ({{r2_train}}) and the testing set ({{r2}}), we observe that the training R^2 is slightly lower than the testing R^2. This suggests that the model is not significantly overfitting the training data. However, the overall R^2 values are relatively low, indicating that the linear regression model might be underfitting the data or that a linear relationship is not the best fit for this dataset.

Impact of Feature Selection

Experimenting with a subset of features ('Pclass' and 'Fare') resulted in a lower R^2 score ({{r2_subset}}) compared to using all features ({{r2}}). This indicates that removing other features in this case increased the bias of the model, making it less capable of explaining the variance in the target variable. The simpler model likely has lower variance but at the cost of increased bias.

Recommendations for Improvement or Next Steps

Based on the analysis, here are some recommendations for improving the model:

1. **Try other regression models:** Linear regression assumes a linear relationship between features and the target variable. Exploring other regression algorithms like polynomial regression, support vector regression, or tree-based models (e.g., Random Forest Regressor, Gradient Boosting Regressor) might yield better results if the relationship is non-linear.

2. **Feature Engineering:** Create new features from existing ones that might capture more complex relationships (e.g., interaction terms, polynomial features).

3. **More advanced feature selection:** Utilize more sophisticated feature selection techniques (e.g., recursive feature elimination, L1 regularization) to identify the most relevant features and potentially improve the bias-variance trade-off.

4. **Hyperparameter tuning:** Optimize the hyperparameters of the chosen model using techniques like cross-validation and grid search to find the best model configuration.

5. **Outlier detection and handling:** Investigate and handle outliers in the data, as they can significantly impact the performance of linear models.

6. **Explore classification models:** Since the target variable 'Survived' is binary (0 or 1), treating this as a classification problem using models like Logistic Regression, Support Vector Machines, or tree-based classifiers might be more appropriate and yield better predictive performance.