Feature Engineering and Data Splitting

Overview

Data preparation is the most basic notion in any dataset to devise predictive models. The biggest tasks done by the feature engineering process are derived via transforming raw information into significant forms that will improve the performance of the model to a great extent. Data splitting ensures that the model is trained and tested on a set of different data subsets to prove the effectiveness and prevent overfitting.

Goals

- To engineer pertinent features that add to the prediction power of the model.
- To divide the dataset into training and testing sets so that the model can see one part and the evaluation can be done on the other to get objective performance metrics.
- To document the whole process toward full reproducibility and transparency.

Steps

1. **Data Understanding**
   - **Objective**: Understand the dataset and which features will be used for the predictive model.
   - **Actions**:
     - Looked at the dataset to have a closer look at categorical and numerical features.
     - Distributions of features and the relation with the target variable (Churn).
2. **Feature Engineering**
   - **Target**: The intent is to make raw features more informative toward the model.
   - **Actions**:
     - Categorical encoding: Used techniques like one-hot encoding and label encoding, converting all the categorical variables into a form understood by the model for training.
     - Feature Scaling: Normalized the number features into a single scale using both Min-Max Scaling and Standard Scaling.
     - Interaction Features: Interaction terms between features, such as Tenure x MonthlyCharges, were created to add information on relationships besides linearity.
     - Derived Features: Created new features from existing ones; for example, get the tenure range of customers.
     - Feature Selection: Selected features after checking the importance scores from the first set of models and measuring the correlation with the target variable.
3. **Data Splitting**
   - **Objective**: Splitting the data into training and testing sets for training and testing the model.
   - **Actions**:
     - Taking into account the target variable's class balance, the dataset's training and test splits were stratified at an 80:20 ratio.
     - Validation Split: To fine-tune the model, the train data are divided 80:20 into training and validation.

- Cross-Validation: During training, K-Fold cross-validation was used to make sure the model is strong and will work well in general.

Importance of Feature Engineering and Data Splitting

- **Feature Engineering**: As it gives the model access to more relevant and instructive data, this is one of the crucial steps in enhancing model performance. Since it makes patterns and links in the data that may not be immediately apparent when viewed directly possible, it can result in significant performance gains.
- **Data Splitting**: Splitting data prevents overfitting and provides a realistic measurement of the model's performance on unseen data. We are able to estimate the model's effectiveness in actual use because this ensures that it is thoroughly evaluated on a different set of data.

Result of Feature Engineering: Developed good features, which resulted in improved performance of the model, confirmed by high accuracy, and good feature importance scores.

Data Split: Build robust test, validation, and training datasets, resulting in a model that is highly adaptable to new data.

Next Steps in Data Validity

- Run the data validity test to ensure that data used for feature engineering and clustering is correct, consistent, and free from errors. This will also entail the check on integrity of the data set and validating if the transformations and splits are properly done.
- Check Model: Verify model performance on the testing set. Refine features or model if necessary.
- Document: Complete the document and update Confluence with the outcomes and insights from the stage of model training.

Conclusion

Feature engineering and splitting of the data are the elementary steps of data preparation. The performance of the model increased due to the engineered features, and a data-splitting strategy created a strong framework of evaluation. In the next steps, the obtained results will focus on the modeling to refine it.