Questions

1. **How does the decision tree-based model address the population distribution of actual yields better than linear regression?**

Decision tree-based models, with their ability to capture non-linear relationships and handle complex distributions, are more adept at accommodating such variations compared to linear regression, which assumes a linear relationship between features and the dependent variable.

2. **What were the key hyperparameters of the Random Forest Regression model, and how did they contribute to its success in predicting bacon yields?**

The key hyperparameters of the Random Forest Regression model are "max_depth," "max_features," "min_sample_leaf," "min_samples_split," and "n_estimators." These hyperparameters contribute to the model's success by controlling the complexity and behavior of the decision trees within the random forest. For example, "max_depth" limits the depth of the individual trees, preventing overfitting.

3. **Can you provide more details on the data preprocessing steps, especially regarding the cleaning and merging of the assembly masters and purchase order datasets?**

Data preprocessing involved stripping header information, converting to lowercase, dropping non-essential columns, and rearranging assembly masters to ensure identical features for concatenation. Purchase order data underwent similar preprocessing, with additional steps like calculating differences and standard deviations, followed by an inner join with assembly masters. Dummy variables for categorical columns were created, and numerical values were converted, leading to a usable dataset for modeling.

4. **What specific features, aside from slice width, showed notable correlations with actual yields?**

The next two highest correlations were “minimum_length_secondary_lean(in)” and “package_type_vacuum_pack’. Their correlations were -0.42 and -0.49 respectively.

5. **How was the data for histocal yields acquired?**

The histocal yields were acquired from an SAP database containing purchase order information from two U.S. based bacon plants. The data is a mix of automated and manually entered data throughout the production process. The yields are acquired by taking the final production pounds of number one products and dividing it by the base materials used.

6. **How large of concern is human error in data?**

A significant concern in the context of outliers is the potential impact on results. Minor errors would not significantly affect the outcomes, but the addition or subtraction of zeros to the total amount of base material received could have a notable effect on the historical yield. This is a key reason for removing anything beyond two standard deviations.

7. **What challenges were encountered during the project, and how were they addressed to ensure the reliability of the predictive model?**

The largest challenge was the small data sets of products available. This was mitigated by decreasing the training and test ratio to 70/30% and focusing on model parameters to reduce overfitting.

8. **How will the developed Jupyter notebook be integrated into the workflow of the engineering standard team?**

The Jupyter notebook will be integrated into the engineering standard team's workflow by using an Excel file with standardized data. Engineers can execute the notebook, placed in a designated folder on the shared drive, without the need for a standalone application, minimizing development time.

9. **Are there any limitations or potential areas of improvement identified in the predictive model that should be considered in future iterations?**

A huge area of improvement would be to create and perform yield test to acquire more accurate yield information of products. Currently each plant grades number one products as they see fit, so it can be a challenge to compare product yields between the two plants.

10. **How much can the predictive model contribute to improving cost predictions, considering the various factors involved in the costing of bacon products?**

The predictive model can substantially contribute to improving cost predictions for bacon products. The largest cost per pound of bacon products made comes from the base material used. So, being able to accurately predict the amount needed can drastically reduce errors in the costing of products.