Bacon: Predicting New Bacon Product Yields

Introduction

The objective of this project is to predict the number one yields on newly created ready to cook (RTC) bacon products. The number one yield is the amount of bacon from a belly that can be used for a primary product, everything else is sold as a secondary cheaper product. The secondary products include low grade slices, side strap trimmings, and ends and pieces. This is important to understand and predict because it affects the costing of the product. Costing allows marketing and sales to accurately predict margin and sell prices of bacon products. Which is vital for a company to succeed in the low margin business model of food production, where a few percent can be the difference between profit and loss.

Currently the yields of in production products are estimated off historical data (12-month increments), but only sporadically updated. While new products are estimated based on similar product's yield. This method does not have a clear standard operating procedure (SOP) and relies on the personnel's knowledge of the specific products. This method can lead to relatively large errors in yield predictions and costing of products. By using quality assurance (QA) information, product specifications, and types of equipment this project will attempt to create a model that can more accurately predict these yields.

Data

The datasets for this project were obtained from two bacon product plants in the United States. Bacon specifications, QA requirements, and equipment types were collected from assembly masters used by the Engineering Standards team. Assembly masters are models/equations that calculate plant throughput (lbs./hr.) and contain essential data about product specifications and plant equipment. The second data source, serving as the dependent variable, comprises historical yields for products produced

at the plants. These yields are determined by weighing all primary produced products and dividing by the weight of the original bacon bellies used.

Once the data was loaded into Python data frames, the header information was stripped and converted to lowercase. Non-essential columns for this project were dropped from the datasets. The two assembly masters were cleaned and rearranged to have identical columns/features, allowing their concatenation into a single dataset. Duplicate products were then removed after combining the two assembly masters.

The purchase order (PO) dataset, containing historical yields over the past 12 months, underwent similar preprocessing. The data was stripped, converted to lowercase, and only the columns product code, target yield, and actual yield were retained. New columns were created to calculate the difference between target and actual yields, as well as the standard deviation in yield differences. Products with a standard deviation greater than two were excluded to mitigate potential human errors in manual data entry. The mean of all "actual yields" was calculated for each product, serving as the dependent variable for the models. Subsequently, the PO data frame was merged with the ASM data frame using an inner join. The new data frame underwent cleaning, removing duplicates, replacing similar strings with identical text, and filling in "NaN" values as needed. To make the data usable for modeling, dummy variables were created for categorical columns, and columns with numerical values were converted to numeric data types. Finally, the data was split into training and test sets based on a 70/30 ratio.

Analysis

The initial phase of the analysis focused on gaining insights into the population distribution of actual yields. Figure 1 below illustrates a distinctive population in the center, flanked by two smaller populations on either side of the curve. This observation suggests that

specific product types adhere to either stricter or more lenient grading requirements. While this characteristic doesn't pose a challenge for model creation, it raised concerns about the capabilities of a linear regression model, prompting a preference for decision tree-based models.

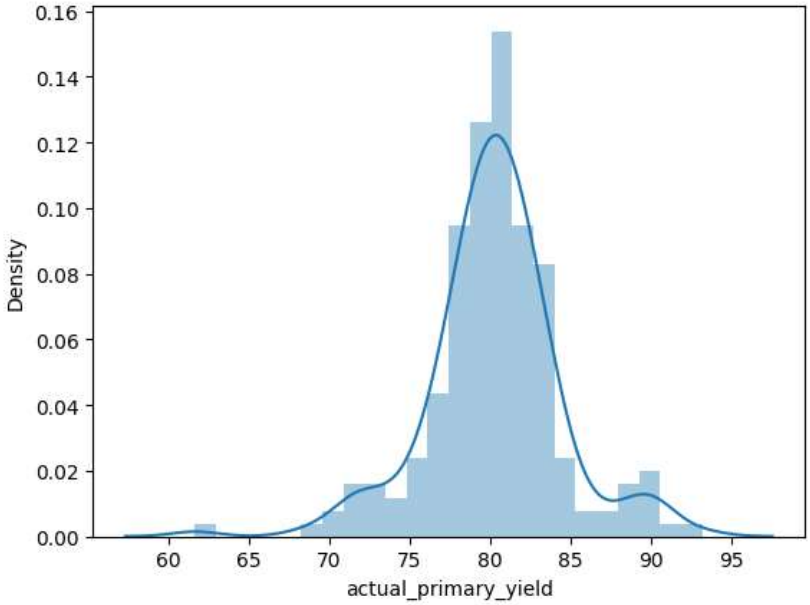

Figure 1: Density Plot of Actual Yield

Afterwards, a correlation heatmap was generated to explore the linear relationships between the features and the dependent variable. The most notable correlation was observed with the "slice_width_upper_(in)" feature, registering a correlation coefficient of 0.62. Other correlations generally fell within the range of absolute values between 0.20 and 0.50, as illustrated in Figure 2. Subsequently, scatterplots were crafted for features exhibiting the highest correlations to provide a visual representation of the relationships. In Figure 3, the scatterplot for the "slice_width_upper_(in)" feature is presented, highlighting a distinct and clear linear relationship. Notably, as the slice width increases, there is a corresponding rise in the average

yield. Which corresponds with the known issue of thinner bacon having more waste due to bacon being damaged during slicing.

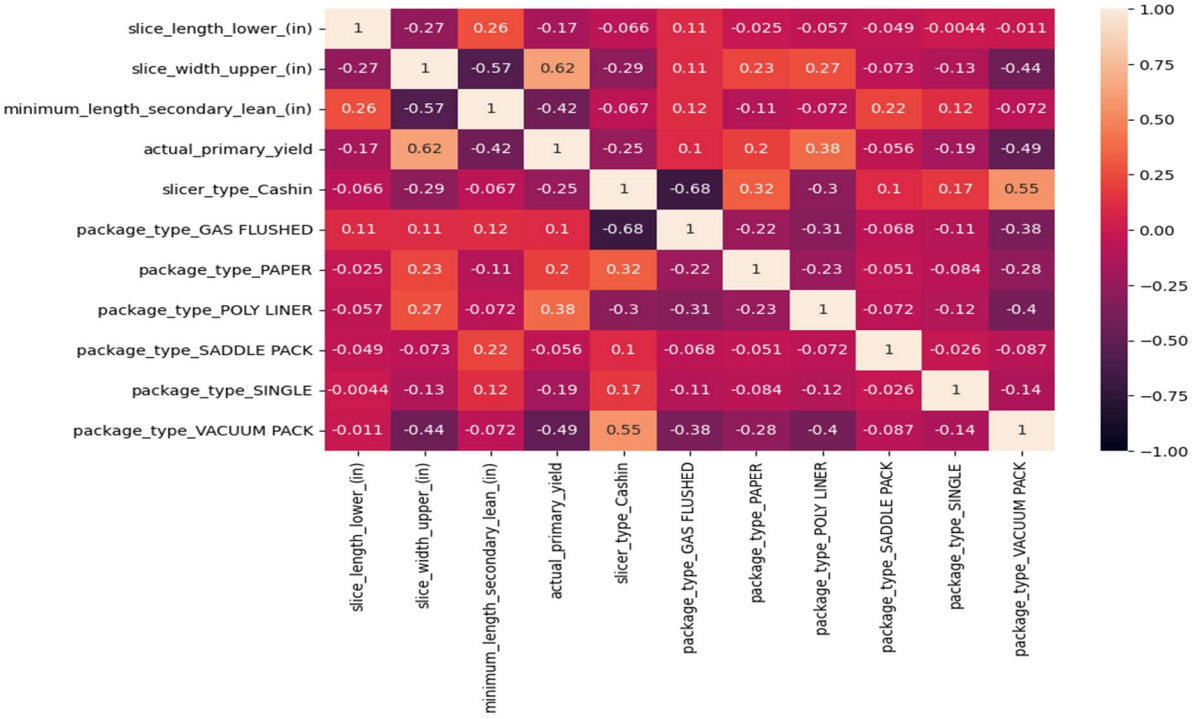

Figure 2: Correlation heatmap

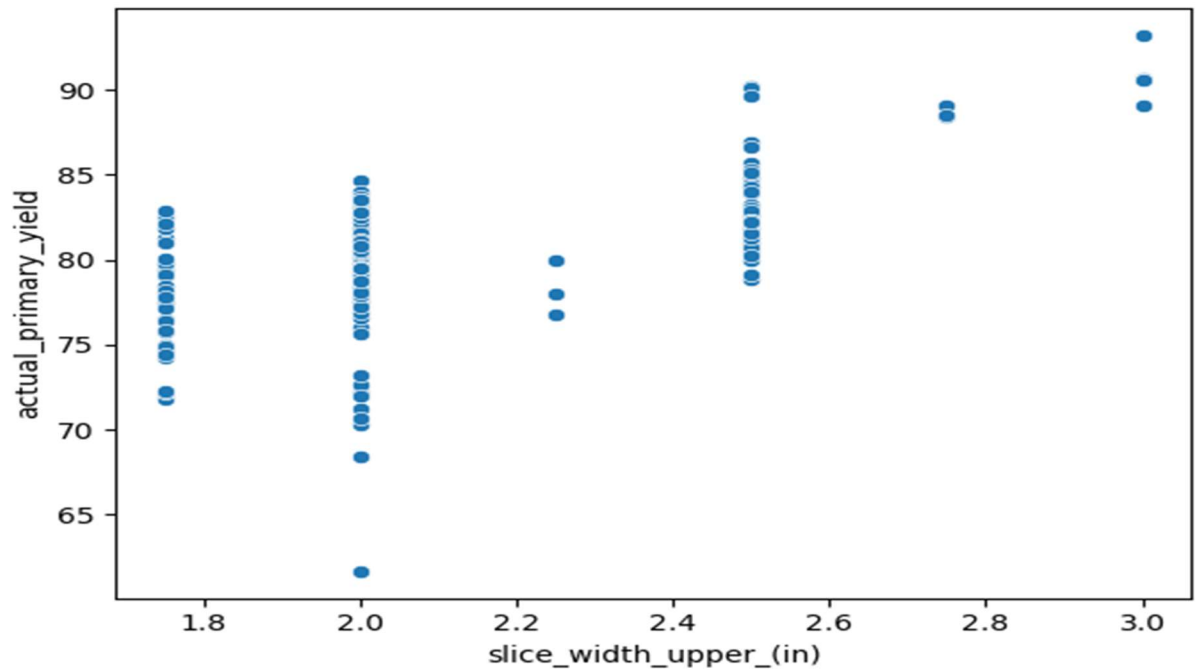

Figure 3: Scatterplot of Actual Yield Vs. Upper Slice Width(in)

Methods

Before constructing a predictive model, the mean absolute error (MAE) and root mean squared error (RMSE) were computed based on the existing method of predicting yields to establish a baseline. This baseline serves as a reference point for assessing the performance improvement achieved by the new model. Additionally, residual plots were generated to facilitate visualizing any enhancements in prediction accuracy.

Following the establishment of the baseline, a pipeline and grid search were implemented utilizing linear regression, random forest, and gradient boost models. Given the relatively small size of the training set, an extensive range of hyperparameters could be tested without significant concerns of computational resources and time constraints. Once the optimal model and hyperparameters were identified, the R-squared, MAE, and RMSE metrics were calculated for both the training and testing datasets to evaluate potential "overfitting" issues. This comprehensive evaluation ensures a thorough assessment of the model's performance and generalization capabilities.

Conclusion

The decision tree-based model demonstrated superior capability in handling the population distribution of the dependent variable, with the Random Forest Regression model identified as the optimal choice among the three tested. The model's hyperparameters, including "max_depth"=6, "max_features"=log2, "min_sample_leaf"=1, "min_samples_split"=2, and "n_estimators"=300, contributed to its overall success, surpassing expectations, and offering a substantial improvement over the existing method.

In contrast to the baseline method, which yielded MAE and RMSE values of 5.18 and 6.49 across all products, the developed model exhibited a significant enhancement. On the test data, the model achieved a MAE of 1.69, RMSE of 2.28, and a R-Squared of 0.708, resulting in a remarkable 67% reduction in prediction errors for yields using only pre-production data. To visually depict the improvement, residual plots (Figure 4 and Figure 5) showcase the decrease in overall errors. It's important to note that these plots are not directly comparable on a one-to-one basis due to differences in datasets used. Specifically, the “current method” plot encompasses 100% of the products, while the “model plot” includes only 70% of the products. Nevertheless, these plots effectively illustrate the considerable enhancement achieved by the model.

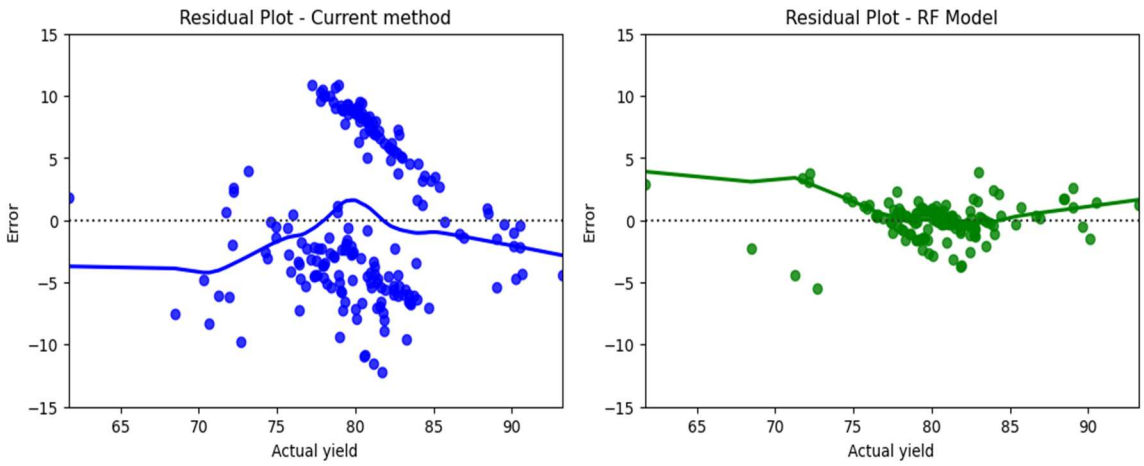

Figure 4 and 5: Residual Plots

The next phase will involve integrating the model into a user-friendly and organized notebook, allowing its use by the engineering standard team for predicting yields of new bacon products. The notebook will be designed to operate with an Excel file containing data for a new bacon product, which should be placed in a designated folder on the shared drive, following a standardized naming convention. This setup enables engineers to efficiently execute the

notebook. Obtaining predicted yields without the necessity of constructing a standalone application, given the specialized and compact nature of the team. Training requirements for utilizing the Jupyter notebook are expected to be minimal.

Ethical Considerations

In general, there are no significant ethical concerns associated with developing and utilizing a model for predicting bacon primary yields. The model predictions cannot be used in any illicit ways and will not have a negative effect on any individuals. The one area of concern would be data security, due to trade secrets. All sensitive information has been removed from the datasets prior to loading. Including all the information that would be needed to perform a full costing of the products. The historical yields by themselves are not enough to calculate the cost of a product. Which includes fixed overheads, corporate overhead, labor, utilities, throughput (Lbs./hr.), volume, packaging material, and base material cost.

References

1. Chugh, A. (2022, March 16). Mae, MSE, RMSE, coefficient of determination, adjusted R squared - which metric is better?. Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
2. Sklearn.ensemble.randomforestregressor. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
3. Brownlee, J. (2021, March 6). XGBoost for regression. MachineLearningMastery.com. https://machinelearningmastery.com/xgboost-for-regression

Appendix

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a metric used to measure the average absolute errors between predicted and actual values. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is another metric for evaluating the accuracy of a model's predictions. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R-Squared

R-Squared, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$