ChocoDelight Factory: Predicting Production Line Issues

Scenario:

Imagine a chocolate bar factory named "ChocoDelight". The factory has a production line consisting of:

1. Cocoa_Roasting_Level: Numeric metric for cocoa bean roast level.
2. Sugar_Granularity: Numeric metric for sugar grain size.
3. Milk_Solid_Percentage: Percentage of milk solid in the mix.
4. Nuts_Inclusion_Rate: Rate of nuts' inclusion in chocolates with nuts.
5. Texture_Smoothness: Numeric metric for chocolate smoothness.
6. Sweetness_Level: Numeric metric for chocolate sweetness.
7. Cocoa_Source_Check: Numeric metric for the source quality of cocoa.
8. Flavor_Infusion: Numeric metric for additional flavors.
9. Shelf_Life_Test: Days the chocolate can stay fresh.
10. Seal_Integrity: Numeric metric for the packaging seal.
11. Barcode_Scan_Success: Rate at which barcodes scan successfully.
12. Box_Dimensions: Numeric metric for packaging box size.
13. Label_Readability: Numeric metric for label clarity.
14. Nutritional_Info_Accuracy: Numeric metric for the accuracy of nutritional information.
15. Chocolate_Glossiness: Numeric metric for chocolate shine.

Certain records are marked as "not okay", indicating a problem at that station. Additionally, there's a dataset that provides **rework time** and **cost** due to issues at each station.

Objective:

Build a model to predict potential production issues, understand the variables leading to these problems, and quantify savings in rework time and costs using causal inference. Display the findings in an interactive dashboard.

Deliverables:

Data Preprocessing:

- Handle missing values.
- Rectify outliers or anomalies.
- Standardize/normalize measurements as needed.

Exploratory Data Analysis (EDA):

- Analyze the distribution of measurements for each station.
- Identify patterns in "not okay" instances.

- Determine correlations between different stations.

Feature Importance Analysis:

- Ascertain which features most influence a station's "not okay" status.
- Use techniques such as recursive feature elimination or permutation importance.
- Visualize top features' relative importance.

Predictive Modeling:

- Develop a model to predict "not okay" situations.
- Assess model metrics.
- Refine model for best performance and focus on Explainability.

Causal Inference:

- Analyze the ‘rework_data’ detailing rework time and costs.
- Simulate scenarios to determine potential savings in rework time and costs.

  Example: If Station XYZ values were higher/ lower, how would the ‘Rework_Time’ and ‘Rework_Cost’ change?

Dashboard Development in PowerBI/ Quicksight or your favourite tool:

Design an interactive dashboard that showcases:

- Predicted vs. actual "not okay" situations.
- Historical measurements for each station.
- Relative importance of top influential features.
- Estimated savings in rework time and costs.