Project Title	**Industrial Copper Modeling**
Skills take away From This Project	**Python scripting, Data Preprocessing, EDA, Streamlit**
Domain	**Manufacturing**

Problem Statement:

The copper industry deals with less complex data related to sales and pricing. However, this data may suffer from issues such as skewness and noisy data, which can affect the accuracy of manual predictions. Dealing with these challenges manually can be time-consuming and may not result in optimal pricing decisions. A machine learning regression model can address these issues by utilizing advanced techniques such as data normalization, feature scaling, and outlier detection, and leveraging algorithms that are robust to skewed and noisy data.

Another area where the copper industry faces challenges is in capturing the leads. A lead classification model is a system for evaluating and classifying leads based on how likely they are to become a customer . You can use the STATUS variable with WON being considered as Success and LOST being considered as Failure and remove data points other than WON, LOST STATUS values.

The solution must include the following steps:

1) Exploring skewness and outliers in the dataset.
2) Transform the data into a suitable format and perform any necessary cleaning and pre-processing steps.
3) ML Regression model which predicts continuous variable ‘Selling_Price’.
4) ML Classification model which predicts Status: WON or LOST.
5) Creating a streamlit page where you can insert each column value and you will get the Selling_Price predicted value or Status(Won/Lost)

Data: Data-set

About the Data:

1. `**id**`: This column likely serves as a unique identifier for each transaction or item, which can be useful for tracking and record-keeping.

2. `**item_date**`: This column represents the date when each transaction or item was recorded or occurred. It's important for tracking the timing of business activities.

3. `**quantity tons**`: This column indicates the quantity of the item in tons, which is essential for inventory management and understanding the volume of products sold or produced.

4. `**customer**`: The "customer" column refers to the name or identifier of the customer who either purchased or ordered the items. It's crucial for maintaining customer relationships and tracking sales.

5. `**country**`: The "country" column specifies the country associated with each customer. This information can be useful for understanding the geographic distribution of customers and may have implications for logistics and international sales.

6. `**status**`: The "status" column likely describes the current status of the transaction or item. This information can be used to track the progress of orders or transactions, such as "Draft" or "Won."

7. `**item type**`: This column categorizes the type or category of the items being sold or produced. Understanding item types is essential for inventory categorization and business reporting.

8. `**application**`: The "application" column defines the specific use or application of the items. This information can help tailor marketing and product development efforts.

9. `**thickness**`: The "thickness" column provides details about the thickness of the items. It's critical when dealing with materials where thickness is a significant factor, such as metals or construction materials.

10. `**width**`: The "width" column specifies the width of the items. It's important for understanding the size and dimensions of the products.

11. `**material_ref**`: This column appears to be a reference or identifier for the material used in the items. It's essential for tracking the source or composition of the products.

12. `**product_ref**`: The "product_ref" column seems to be a reference or identifier for the specific product. This information is useful for identifying and cataloging products in a standardized way.

13. `**delivery date**`: This column records the expected or actual delivery date for each item or transaction. It's crucial for managing logistics and ensuring timely delivery to customers.

14. `**selling_price**`: The "selling_price" column represents the price at which the items are sold. This is a critical factor for revenue generation and profitability analysis.

Approach:

1) Data Understanding: Identify the types of variables (continuous, categorical) and their distributions. Some rubbish values are present in 'Material_Reference' which starts with '00000' value which should be converted into null. Treat reference columns as categorical variables. INDEX may not be useful.
2) Data Preprocessing:
    - Handle missing values with mean/median/mode.
    - Treat Outliers using IQR or Isolation Forest from sklearn library.
    - Identify Skewness in the dataset and treat skewness with appropriate data transformations, such as log transformation(which is best suited to transform target variable-train, predict and then reverse transform it back to original scale eg:dollars), boxcox transformation, or other techniques, to handle high skewness in continuous variables.
    - Encode categorical variables using suitable techniques, such as one-hot encoding, label encoding, or ordinal encoding, based on their nature and relationship with the target variable.
3) EDA: Try visualizing outliers and skewness(before and after treating skewness) using Seaborn's boxplot, distplot, violinplot.
4) Feature Engineering: Engineer new features if applicable, such as aggregating or transforming existing features to create more informative representations of the data. And drop highly correlated columns using SNS HEATMAP.
5) Model Building and Evaluation:
    - Split the dataset into training and testing/validation sets.
    - Train and evaluate different classification models, such as ExtraTreesClassifier, XGBClassifier, or Logistic Regression, using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and AUC curve.

- Optimize model hyperparameters using techniques such as cross-validation and grid search to find the best-performing model.
- Interpret the model results and assess its performance based on the defined problem statement.
- Same steps for Regression modelling.(note: dataset contains more noise and linearity between independent variables so itll perform well only with tree based models)

6) Model GUI: Using streamlit module, create interactive page with

   (1) task input(Regression or Classification) and

   (2) create an input field where you can enter each column value except 'Selling_Price' for regression model and except 'Status' for classification model.

   (3) perform the same feature engineering, scaling factors, log/any transformation steps which you used for training ml model and predict this new data from streamlit and display the output.

7) Tips: Use pickle module to dump and load models such as encoder(onehot/ label/ str.cat.codes /etc), scaling models(standard scaler), ML models. First fit and then transform in separate line and use transform only for unseen data

   Eg: scaler = StandardScaler()

   scaler.fit(X_train)

   scaler.transform(X_train)

   scaler.transform(X_test_new) #unseen data

The learning outcomes of this project are:

1. Developing proficiency in Python programming language and its data analysis libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and Streamlit.
2. Gaining experience in data preprocessing techniques such as handling missing values, outlier detection, and data normalization to prepare data for machine learning modeling.
3. Understanding and visualizing the data using EDA techniques such as boxplots, histograms, and scatter plots.

4. Learning and applying advanced machine learning techniques such as regression and classification to predict continuous and binary target variables, respectively.
5. Building and optimizing machine learning models using appropriate evaluation metrics and techniques such as cross-validation and grid search.
6. Experience in feature engineering techniques to create new informative representations of the data.
7. Developing a web application using the Streamlit module to showcase the machine learning models and make predictions on new data.
8. Understanding the challenges and best practices in the manufacturing domain and how machine learning can help solve them.

Overall, this project will equip you with practical skills and experience in data analysis, machine learning modeling, and creating interactive web applications, and provide you with a solid foundation to tackle real-world problems in the manufacturing domain.

Project Evaluation metrics:

- You are supposed to write a code in a modular fashion (**in functional blocks**)
- Maintainable: It can be maintained, even as your codebase grows.
- Portable: It works the same in every environment (operating system)
- You have to maintain your code on **GitHub**.(Mandatory)
- You have to keep your **GitHub** repo public so that anyone can check your code.(Mandatory)
- Proper readme file you have to maintain for any project development(Mandatory)
- You should include basic workflow and execution of the entire project in the readme file on **GitHub** (Mandatory)
- Follow the coding standards: https://www.python.org/dev/peps/pep-0008/
- You need to Create a Demo video of your working model and post in **LinkedIn(Mandatory)**