Problem Statement: Sculpture Shipping Cost Prediction

You work for a company that specializes in selling sculptures created by artists from around the world. One of the challenges your company faces is accurately predicting the cost required to ship these sculptures to customers. Shipping art pieces, particularly sculptures, involves unique considerations and challenges that differ from shipping regular consumer goods. The goal of this project is to develop a predictive model that estimates the cost of shipping sculptures to their respective destinations based on relevant information provided in the dataset.

Key Points:

- The company sells sculptures from various artists globally.

- The dataset includes a comprehensive set of features to capture the complexities of shipping sculptures:

  - Sculpture dimensions (length, width, height)

  - Sculpture weight

  - Material the sculpture is made of

  - Artist reputation (an indicator of the artist's market standing)

  - Base shipping price

  - Whether the shipping is international

  - Whether the shipping was in express (fast) mode

  - Whether installation is included in the purchase

  - Mode of transport

  - Whether the order is fragile

  - Customer information and location

  - Whether the customer resides in a remote location

  - Scheduled date and delivery date

- Shipping costs for sculptures are not as straightforward as for consumer goods due to their unique nature and the variety of factors involved.

- The predictive model will consider the comprehensive set of features to provide customers with highly accurate shipping cost estimates, thereby enhancing transparency and customer satisfaction.

- The initial focus is on creating a basic Proof of Concept (POC) to demonstrate the feasibility of predicting shipping costs based on the provided dataset.

Objective:

The primary objective of this Proof of Concept (POC) is to design and develop a machine learning model that leverages a comprehensive set of features, including sculpture dimensions, weight, distance, shipping service type, artist reputation, material, and various shipping details. The model will predict the corresponding shipping cost, thus providing customers and stakeholders with accurate and transparent estimates for shipping sculptures. This project aims to showcase the potential of machine learning in addressing the unique challenges of predicting shipping costs for art pieces, thereby enhancing customer satisfaction and optimizing logistical operations.

Key Deliverables:

1. Dataset Preparation: Clean, preprocess, and engineer the dataset to ensure it's suitable for model training. Handle missing values, outliers, and perform necessary transformations.

2. Model Training and Evaluation: Train the model using an appropriate algorithm and evaluate its performance using relevant metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

3. Machine Learning Model: Develop a predictive machine learning model that considers the diverse range of features in the dataset. The model will be trained to estimate the shipping cost of sculptures based on these features.

4. User Interface: Create a user-friendly interface that allows users (customers, company staff, and stakeholders) to input sculpture details and obtain estimated shipping costs. The interface should be intuitive and accessible to provide a seamless experience.

- Incorporating advanced machine learning techniques to capture intricate relationships in the dataset.

- Adding features such as customer preferences, historical shipping data, and external factors affecting shipping costs.

- Developing a more refined user interface with visualizations and user-specific accounts.

Outcome:

By successfully completing this POC, we aim to showcase the effectiveness of machine learning in accurately predicting shipping costs for sculptures. The user interface will provide an accessible way for users to obtain estimates, improving transparency and customer experience.

Step 1: Data Collection and Preparation

In this step, we gather and prepare the dataset for predicting shipping costs for sculptures. The dataset has been obtained from Kaggle and includes information about sculpture characteristics and associated shipping costs.

Data Source -Kaggle dataset - "Shipping Cost Prediction"

Features:

1. Customer Id: Unique identification number of the customers.
2. Artist Name: Name of the artist.
3. Artist Reputation: Reputation of the artist in the market.
4. Height, Width, Weight: Dimensions and weight of the sculpture.
5. Material: Material that the sculpture is made of.
6. Price Of Sculpture: Price of the sculpture.
7. Base Shipping Price: Base price for shipping a sculpture.
8. International: Whether the shipping is international.
9. Express Shipment: Whether the shipping was in the express (fast) mode.
10. Installation Included: Whether the order had installation included.
11. Transport: Mode of transport of the order.
12. Fragile: Whether the order is fragile.
13. Customer Information: Details about a customer.
14. Remote Location: Whether the customer resides in a remote location.
15. Scheduled Date: Date when the order was placed.
16. Delivery Date: Date of delivery of the order.
17. Customer Location: Location of the customer.
18. Cost: Represents the cost of the order (target variable).

Actions:

1. Load the dataset from the provided Kaggle source.

2. Explore the dataset to understand its structure, feature distributions, and relationships between variables.

3. Handle Missing Values: Check for missing values in the dataset and apply appropriate strategies to handle them (e.g., imputation).

4. Outlier Handling: Identify and deal with outliers in relevant features such as dimensions, weight, price, and shipping cost.

5. Convert Categorical Variables: Apply one-hot encoding to convert categorical variables like shipping service type, material, and transport mode.

6. Data Cleaning: Ensure the dataset is cleaned of inconsistencies and errors, and that all features are in suitable formats for analysis.

Outcome:

A cleaned and preprocessed dataset is prepared, containing numerical representations of features, no missing values, and appropriate formats. This dataset will be used for subsequent steps in the POC, including model building and evaluation.

Data Preprocessing Class

Introduction:

Welcome to the Data Preprocessing Class! In this class, we'll focus on preparing the dataset for analysis and modeling. You'll find user-defined functions for handling missing values, outliers, and feature engineering.

Class Structure:

1. Data Loading and Exploration:

   - Load the dataset into the class.

- Explore the dataset's structure, features, and relationships between variables.

2. Missing Value Handling:

- Function: FillMissingNumericalVariables:

Define and apply strategies for filling missing values in numerical variables. Options could include mean, median, or custom values.

- Function: FillMissingCategoricalVariables:

Define and apply strategies for filling missing values in categorical variables. Options could include mode, creating a new category, or custom values.

3. Outlier Handling:

- Function: HandleOutliers:

Define methods for identifying and handling outliers in specific numerical variables. Options could include capping, transformations, or removal.

4. Feature Engineering:

- Function: CreateNewFeatures:

Define and implement new feature creation methods. This could involve combining or transforming existing features to extract valuable insights.

5. Data Cleaning and Transformation:

- Function: DataCleaning:

Define specific tasks to clean and transform the dataset, ensuring it's consistent and suitable for analysis.

6. Conclusion:

By utilizing the functions provided in this class, you'll be equipped with essential tools to preprocess data effectively, making it ready for analysis and modeling.

Data Transformation Pipeline Class

Introduction:

Welcome to the Data Transformation Pipeline Class! In this class, we'll delve into advanced data transformations, including pipeline creation for both categorical and numerical data. You'll define and implement pipelines to streamline your data preparation process.

Class Structure:

1. Data Loading and Exploration:

- Load the dataset into the class.
- Explore the dataset's structure, features, and relationships between variables.

2. Data Transformation Pipelines:

Categorical Data Pipeline:

- Function: CategoricalPipeline:

Define a pipeline to handle categorical data transformations, including missing value handling and feature engineering specific to categorical features.

Numerical Data Pipeline:

- Function: NumericalPipeline:

Define a pipeline to handle numerical data transformations, including missing value handling, outlier handling, and feature engineering specific to numerical features.

Combined Preprocessing Pipeline:

- Function: CombinePipelines:

Combine the categorical and numerical pipelines into a comprehensive preprocessor.

3. Conclusion:

By mastering the functions and pipelines provided in this class, you'll be able to efficiently transform and preprocess diverse datasets, ensuring they are optimized for analysis and modeling tasks.

Step -2 Model training & Evaluation

Welcome to Step 2 of our journey: Model Training & Evaluation! In this phase, we'll dive into the exciting process of training machine learning models and assessing their performance. As we explore this crucial step, we'll learn how to select appropriate models, train them using the prepared data, and evaluate their effectiveness in making predictions. Let's get started!

Introduction:

In this phase, we'll focus on the heart of predictive analytics: building models that can understand patterns in the data and make accurate predictions. But before we jump into model training, let's remind ourselves of the broader goal. We're on a mission to predict the shipping costs of sculptures accurately. To achieve this, we'll leverage the preprocessed dataset we crafted in the previous steps.

Key Objectives:

1. Model Selection: We'll begin by understanding the available machine learning algorithms and selecting the most suitable ones for our problem. The choice of model depends on the nature of our dataset, the complexity of relationships, and the prediction task.

2. Feature Importance: We'll explore which features have the most significant impact on predicting shipping costs. This insight can guide us in refining the model's input and potentially improving prediction accuracy.

3. Training and Tuning: With our chosen models, we'll split the dataset into training and validation sets. We'll then train the models on the training data and fine-tune their hyperparameters to optimize performance.

4. Evaluation Metrics: To assess how well our models are performing, we need reliable evaluation metrics. We'll discuss and employ metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to measure the accuracy of our predictions.

5. Model Comparison: We'll compare the performance of different models to identify the best-performing one. This comparison will guide our selection of the final model to be used in the shipping cost prediction process.

Outcome:

By the end of this step, you'll have gained a deeper understanding of model training and evaluation. You'll be equipped with the knowledge and skills to make informed decisions when it comes to selecting the right machine learning algorithms, fine-tuning their parameters, and quantifying their performance. The journey to accurate shipping cost predictions is well underway!