**FUTURE SALES PREDICTIONS**

**TEAM MEMBERS**

**PRIYANKA PANDA(**[**priyankapanda127@gmail.com**](mailto:priyankapanda127@gmail.com)**)**

**SHREYA CHOUDHARY (** [**shreyachoudhary1903@gmail.com**](mailto:shreyachoudhary1903@gmail.com)**)**

**KARISHMA KAINE(**[**21l121@psgitech.ac.in**](mailto:21l121@psgitech.ac.in)**)**

**DHANYA SHREE M (dhanyashreemanikam@gmail.com)**

**PRITHISH GOUTAM S (**[**kanthaex@gmail.com**](mailto:kanthaex@gmail.com)**)**

**SANKAR MALAI MOHAN PS (**[**d22l602@psgitech.ac.in**](mailto:d22l602@psgitech.ac.in)**)**

**PROBLEM STATEMENT FOR FUTURE SALES PREDICTION:**

"The problem at hand is to create a forecasting model that can reliably estimate sales for a future time frame, leveraging historical sales data and relevant variables. The aim is to enhance business planning and decision-making by providing accurate sales projections for a specified future period."

**DESIGN THINKING PROCESS**

**NEEDS:**

1. Accurate Data: High-quality, up-to-date data is essential for making accurate predictions. This may involve gathering data from various sources, including sensors, surveys, and online platforms.

2. Advanced Models: Developing and utilizing advanced predictive models, such as machine learning algorithms, statistical methods, and simulation models, is crucial for improving prediction accuracy.

3. Interdisciplinary Approach: Future-scale predictions often require collaboration between experts in different fields, such as data science, economics, and domain-specific knowledge, to provide a holistic view.

**GOALS:**

1. Forecasting: The primary goal is to accurately forecast future scales, such as population growth, economic trends, climate change impacts, or disease outbreaks.

2. Risk Assessment: Identify potential risks associated with changes in scales and develop strategies to mitigate these risks.

3. Resource Allocation: Assist in the allocation of resources, whether for infrastructure development, healthcare, or disaster preparedness, to meet the changing scale demands.

4. Policy Formulation: Provide insights for policymakers to create informed policies and regulations that address the challenges and opportunities associated with changing scales.

5. Continuous Improvement: Continuously refine prediction models and methods as new data becomes available and technology advances.

**BRIEF VIEW OF THE PROBLEM STATEMENT**

**Data Availability:** The problem involves utilizing historical sales data, which may include information about sales by product, region, time period, and other relevant dimensions. Ensure data availability, quality, and completeness to construct a reliable forecasting model.

**Time Frame:** Specify the future time frame for which sales forecasting is required. This could be monthly, quarterly, annually, or any other relevant period.

**Accuracy and Reliability:** The goal is to create a forecasting model that is highly accurate and reliable. The model should minimize errors in sales projections to enable effective business planning.

**Variable Consideration:** Identify and incorporate relevant variables and factors that influence sales. This might include factors like marketing campaigns, economic indicators, seasonality, and competitive forces.

**Business Planning**: The primary objective is to support business planning and decision-making. Accurate sales projections enable organizations to allocate resources effectively, plan inventory, manage staff, and optimize marketing strategies.

**Model Selection:** Determine the appropriate forecasting method or model to use. Options might include time series analysis, regression analysis, machine learning, or a combination of techniques.

**Evaluating Model Performance:** Establish criteria for evaluating the model's performance, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). The model should be validated against historical data to ensure its effectiveness.

**Technology and Tools:** Determine the technology stack and tools that will be used for data preprocessing, model development, and visualization.

**SOLUTION:**

**Machine Learning and AI Integration**: Utilize machine learning and artificial intelligence to create predictive models that can adapt to evolving data and complex interdependencies in scale.

**Predictive Analytics:** Establish centralized predictive analytics that pool resources and expertise to tackle large-scale prediction challenges collaboratively.

**MOCK SCALES PREDICTION MODELS**

**Scenario:** Predicting Future Urban Population Growth

**Model Components:**

**Linear Regression Model**: A simple linear regression model can predict population growth based on historical population data and economic indicators. The model would provide a basic projection of population growth if current trends continue.

**Assumptions and Limitations:**

- These models rely on historical data and make assumptions about future trends. Unexpected events can disrupt predictions.

- Accuracy depends on data quality and the complexity of the model used.

- Social and cultural factors may not be fully captured by these models.

**DEVELOPMENT OF THE SCALE PREDICTION SOLUTION**

1. Data Gathering and Preparation:

- Collect a comprehensive dataset of historical sales data, including factors that influence sales.

- Clean and preprocess the data, handling missing values and outliers.

2. Feature Engineering:

- Create relevant features from the data that can improve prediction accuracy, such as lag variables, moving averages, or interaction terms.

3. Model Selection:

- Choose an appropriate machine learning model for sales prediction. Options may include linear regression, time series models (ARIMA, Prophet), or more advanced techniques like neural networks.

4. Training and Validation:

- Split the data into training and validation sets. Use the training set to train the model and the validation set to fine-tune hyperparameters.

5. Model Evaluation:

- Assess the model's performance using appropriate metrics (e.g., Mean Absolute Error, Root Mean Squared Error) on the validation set.

6. Feature Importance Analysis:

- Analyze feature importance to understand which variables have the most significant impact on sales.

**PHASES OF DEVELOPMENT**

1. Project Initiation:

The objective of the future scales prediction project is to develop an accurate and reliable forecasting model that leverages historical data and relevant variables to provide precise sales projections for a specified future time frame, ultimately enabling informed business planning and decision-making.

2. Data Collection and Preparation:

Gather historical data relevant to the scales you want to predict. This data can come from various sources, such as databases, surveys, sensors, or online sources.

Clean and preprocess the data to handle missing values, outliers, and inconsistencies.

Explore and analyze the data to gain insights into the scales' historical trends, patterns, and relationships.

3. Feature Engineering:

Identify relevant features (variables) that can influence the scales you're predicting. This may involve domain knowledge and data analysis.

Create new features or transform existing ones to improve model performance.

Normalize or standardize features as needed.

4. Model Selection:

We have chosen linear regression model.

Linear regression is a commonly chosen model for future sales prediction for several reasons:

1. Simplicity: Linear regression is a simple and easy-to-understand modelling technique. It assumes a linear relationship between the independent variables (features) and the dependent variable (sales), making it accessible to a wide range of users.

2. Interpretability: The coefficients in a linear regression model provide direct insights into how each independent variable influences sales. This interpretability can be crucial for businesses trying to understand the drivers of their sales.

3. Speed and efficiency: Linear regression is computationally efficient and can be trained and used for prediction relatively quickly, which is important for real-time or near-real-time sales forecasting.

4. Baseline model: Linear regression can serve as a baseline model for sales prediction. While more complex models may provide better accuracy in some cases, linear regression provides a reasonable starting point, and more sophisticated models can be compared to it to assess their performance.

5. Robustness: Linear regression can be robust to outliers if appropriate data preprocessing techniques are applied, such as data transformation or outlier handling.

5. Model Training:

Divide the data into training, validation, and test sets for model development and evaluation.

Train the chosen models on the training data.

Evaluate models using the validation set, and iteratively refine them based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or others relevant to the project.

6. Model Evaluation:

Assess the model's predictive accuracy and generalization by evaluating it on the test dataset.

Utilize various evaluation metrics to determine how well the model performs. Common metrics include R-squared, Root Mean Squared Error (RMSE), or classification metrics if predicting categorical scales.

Conduct cross-validation to ensure robustness and reduce overfitting.

7. Deployment:

Once a satisfactory model is developed, deploy it into a production environment. This may involve integrating the prediction model into an application, website, or other systems.

Implement monitoring and maintenance strategies to ensure the model continues to perform well as new data becomes available.

8. Interpretability and Communication:

Ensure that the model's predictions are interpretable and understandable to stakeholders, especially if they impact decision-making.

Communicate results and insights to relevant stakeholders through reports, dashboards, or presentations.

**DESCRIPTION FOR THE DATA SET**

**Link-**<https://www.kaggle.com/datasets/chakradharmattapalli/future-sales-prediction>

The dataset link mentioned above is related to sales prediction and includes information about different advertising mediums and their impact on sales. Here's what each column likely represents:

1. **TV:** This column likely represents the advertising budget spent on TV advertising for a specific time period.

2. **Radio:** This column likely represents the advertising budget spent on radio advertising for the same time period.

3**. Newspaper:** This column likely represents the advertising budget spent on newspaper advertising.

4. **Sales**: This column likely represents the actual sales figures for the product or service during the same time period.

With this information, it's understandable that this dataset is used to explore the relationship between advertising expenditures on TV, radio, and newspaper and the resulting sales figures. This type of dataset is often used in marketing and data analysis to build predictive models or understand which advertising channels have the most impact on sales.

To perform a more detailed analysis or build predictive models using this dataset, we are considering tasks such as regression analysis to predict sales based on advertising spending or exploring the correlation between different advertising mediums and sales performance.

**DATA PREPROCESSING STEPS**

**1. Data Cleaning:**

- Handle missing values: Decide whether to impute missing data or remove rows/columns with missing values.

- Detect and handle duplicates.

- Identify and handle outliers that can adversely affect your analysis.

**2. Data Exploration and Visualization:**

- Visualize the data with histograms, scatter plots, or other relevant plots to

understand the distribution of variables and relationships between variables.

- Explore correlations between features using correlation matrices or scatter plots.

**3.Data Splitting: Training and Testing Sets**

Purpose: The data splitting process is a crucial step in preparing your dataset for machine learning model development and evaluation. It enables you to assess your model's ability to generalize to new, unseen data by reserving a portion of the dataset for testing, while the rest is used for training the model.

**Steps:**

1. Import Necessary Libraries: Begin by importing the necessary libraries and modules for data manipulation and modelling. Common libraries include NumPy, Pandas, and sci-kit-learn.

2. Load and Prepare Your Data: Load your dataset and perform any required data preprocessing steps. Ensure that your data is clean and that both the target variable and features are properly formatted.

3. Decide on the Split Ratio: Determine the ratio in which you want to divide your data into training and testing sets. The standard split is 70-80% for training and 20-30% for testing, but this can vary based on dataset size and problem requirements.

4. Split the Data: Use a data splitting function or method provided by your programming environment or machine learning library to create training and testing sets. For example, in sci-kit-learn, the `train\_test\_split` function is commonly used.

- `X\_train`: This variable contains the feature data for training.

- `X\_test`: This variable contains the feature data for testing.

- `y\_train`: This variable holds the target variable for training.

- `y\_test`: This variable holds the target variable for testing.

5. Confirm the Split: After performing the data split, verify that the sizes of the training and testing sets meet your expectations. Utilize the `shape` or `len` functions to check the number of data points in each set.

6. Proceed with Model Building: With the training data in hand, continue with the development and training of your machine learning model. The testing data is reserved for evaluating the model's performance using appropriate metrics relevant to your specific problem.

**MODEL TRAINING PROCESS**

The dataset, which includes advertising expenditures (TV, radio, newspaper) and sales, can be used to build predictive models to understand how advertising spending affects sales. Below are a few modelling approaches you can consider:

1. Linear Regression:

- Linear regression is a simple yet powerful model for predicting a numerical outcome based on one or more predictor variables (TV, radio, newspaper).

- You can build a linear regression model to quantify the relationship between advertising budgets and sales. The coefficients of the model will show the impact of each medium on sales.

2. Multiple Linear Regression:

- This is an extension of linear regression, which allows you to model the relationship between multiple predictor variables (TV, radio, newspaper) and the target variable (sales).

- By using multiple linear regression, you can account for the combined effect of all advertising channels on sales.

Linear regression is a straightforward and interpretable model used for predicting a continuous target variable based on one or more predictor variables. Here's a more detailed breakdown of the model training process specifically using linear regression:

1. Data Preparation:

- Ensure your dataset is cleaned, preprocessed, and split into training and testing sets, as described in previous responses.

2. Select Linear Regression:

- Since you've chosen linear regression, initialize a linear regression model.

```python

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

```

3. Train the Model:

- Fit the linear regression model to the training data using the training features (X\_train) and the corresponding target variable (y\_train). The model learns the coefficients for each feature to make predictions.

4. Model Evaluation:

- If you have a testing set, evaluate the linear regression model's performance. Common metrics for regression include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to assess the model's goodness of fit.

5. Interpretation:

- Linear regression models provide interpretable coefficients for each feature. You can interpret these coefficients to understand the impact of each feature on the target variable.

- Linear regression has few hyperparameters to tune, but you can explore regularized versions such as Lasso or Ridge regression if needed.

6. Finalize the Model:

- Once you are satisfied with the model's performance, finalize the training process and save the linear regression model for future use.

7.Monitoring and Maintenance :

- Continuously monitor the model's performance in production and retrain it periodically to maintain its accuracy as data evolves.

**Time Series Forecasting Algorithm:**

1. STL (Seasonal Decomposition of Time Series): Decompose the data to understand its components.

2. ARIMA (Autoregressive Integrated Moving Average): Effective for capturing autocorrelation, trend, and seasonality.

3. Prophet: User-friendly tool for daily seasonality and events.

4. Exponential Smoothing (ETS): Suitable for various patterns and changing dynamics.

5. LSTM or GRU (Deep Learning): For complex, non-linear data with ample samples.

6. XGBoost or LightGBM: Useful for structured time series data.

**Evaluation Metrics:**

1. MAE (Mean Absolute Error): Average absolute difference between actual and predicted sales.

2. MSE (Mean Squared Error): Square of MAE, giving more weight to larger errors.

3. RMSE (Root Mean Squared Error): Interpretable measure in the same unit as sales data.

4. MAPE (Mean Absolute Percentage Error): Percentage error, useful for relative error assessment.

5. R² (R-squared): Measures how much variance the model explains.

6. AIC and BIC: Compare models based on lower values for better fit.

7. Forecast Accuracy at Different Time Horizons: Assess accuracy for various prediction periods.