Zewail City of Science and Technology

مدينة زويل للعلوم والتكنولوجيا

Project Name: Sales Forecasting and Demand Prediction

Name	ID
Reem Ehab Helmy	202201373
Tasneem Ashraf Elsaadany	202201573
Mariam Mohamed Goda	202201223

Problem Statement and Importance

Accurate demand and sales forecasting are essential for improving inventory control, cutting expenses, and raising customer satisfaction in the retail industry. Inefficiencies like overstocking or stockouts can result from traditional forecasting methods' inability to account for complicated variables like weather patterns, economic indicators, and changing consumer behavior. By examining huge datasets to find trends and raise the precision of demand forecasts, the application of sophisticated machine learning (ML) algorithms can overcome these difficulties. In dynamic marketplaces where staying competitive requires quick adjustments to consumer preferences, this strategy is especially crucial.

Proposed Algorithms/Models

The following machine learning models and algorithms are suggested to improve demand prediction and sales forecasting:

1. Artificial Neural Networks (ANN): ANNs may learn from past sales patterns and use a variety of influencing elements to increase prediction accuracy. They are also capable of modeling complicated nonlinear interactions inside data.
2. Random Forests (RF): An ensemble learning technique that improves prediction accuracy by building several decision trees and combining their results, RFs work well with big datasets that contain a lot of variables.
3. Support Vector Machines (SVM): SVMs are useful in high-dimensional spaces because they can accurately estimate demand by modeling intricate correlations in sales data.
4. Extreme Learning Machines (ELMs): Distinguished by their quick training times and robust generalization skills, ELMs are capable of processing massive volumes of sales data effectively for forecasting.
5. Ensemble Learning : According to the research papers , the ensemble learning technique used to combine multiple models with each other to improve the accuracy of the model and also to enhance the predictive data . As the abstract mentioned, it can implement Gradient Boosting Machines to treat the imbalance data and the voting classification to combine the predictive data from each model and enhance the accuracy.

6. SHAP (SHapley Additive exPlanations) : it is an algorithm that works on enhancing the explainability as it is an art method used for explaining the machine learning models output.
7. Clustering techniques : clustering is going to use it as it has the ability to group similar data points together as it is going to help to segment customers into groups based on their behavior as the abstract of the paper mentioned.
8. Data Mining models : as the features selection algorithm has the ability to Identify the most relevant features for churn prediction.
9. Seasonal ARIMA (SARIMA): A variation of the ARIMA model that adds seasonal autoregressive, differencing, and moving average components to account for seasonality. SARIMA works especially well with time series data that exhibits recurring seasonal patterns, which makes it a good fit for forecasting sales and predicting demand in cyclically influenced industries like retail and tourism.
10. Extreme Gradient Boosting (XGBoost): A powerful gradient boosting algorithm that optimizes decision trees with advanced regularization, enhancing prediction accuracy. Its efficiency with large datasets and ability to handle complex variable interactions make it a top choice for demand prediction and sales forecasting across industries.

Used Data Set and its description

The data that we are going to use is Walmart Recruiting - Store Sales Forecasting that is located on the kaggle platform .The dataset is provided by Walmart as part of a recruiting competition hosted on Kaggle. you can find it on this Link.

It is a csv files consisting of three files which are :

- Train file: that contains the feathers that models are going to train on it.
- Test file : that contains the data that is needed for the prediction.
- Feature file : and this one contains additional features for training.
- Store file : this file contains some information about the stores.

How will we get it ?

- First we need to combine the three files as a preprocessing stage to enhance the ability of working with the features.
- Also encode the categorical columns also handling the missing data and file it using the sklearn algorithms for it.
- Then start the feature engineering stage and train the used model to start the prediction.

References

- **Research Paper_1: Reem**

1. https://link.springer.com/chapter/10.1007/978-3-319-33681-7_40
2. https://ieeexplore.ieee.org/abstract/document/8659115
3. https://ieeexplore.ieee.org/abstract/document/9117395
4. https://www.mdpi.com/2227-7390/10/17/3179
5. https://hal.science/hal-01881362/

- **Research Paper_2: Mariam**

6. https://ieeexplore.ieee.org/abstract/document/9121220
7. https://www.sciencedirect.com/science/article/abs/pii/S0957417416306327
8. https://pubsonline.informs.org/doi/abs/10.1287/ijoc.2021.1147

- **Research Paper_3: Tasneem Ashraf**

9. https://link.springer.com/chapter/10.1007/978-3-030-28377-3_27
10. https://link.springer.com/chapter/10.1007/978-3-030-36718-3_39

11. https://onlinelibrary.wiley.com/doi/full/10.1155/2019/8503252
12. https://www.sciencedirect.com/science/article/pii/S1877050922003076