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

Used Research Papers(Submitted with new one):

- **Research Paper_1: Reem**
  1. https://ieeexplore.ieee.org/abstract/document/8659115
  2. https://ieeexplore.ieee.org/abstract/document/9117395
  3. https://www.mdpi.com/2227-7390/10/17/3179
  4. https://link.springer.com/chapter/10.1007/978-981-15-0199-9_11
- **Research Paper_2: Mariam**
  1. https://ieeexplore.ieee.org/abstract/document/9121220
  2. https://www.sciencedirect.com/science/article/abs/pii/S0957417416306327
  3. https://www.sciencedirect.com/science/article/abs/pii/S0957417408005800
  4. https://link.springer.com/article/10.1007/s00477-018-1638-6
- **Research Paper_3: Tasneem Ashraf**

1. https://link.springer.com/chapter/10.1007/978-3-030-28377-3_27
2. https://link.springer.com/chapter/10.1007/978-3-030-36718-3_39
3. https://onlinelibrary.wiley.com/doi/full/10.1155/2019/8503252
4. https://www.sciencedirect.com/science/article/pii/S1877050922003076

Literature Review for each one :

- **(Intelligent Sales Prediction Using Machine Learning Techniques)_1_Reem :-**

  The paper "Intelligent Sales Prediction Using Machine Learning Techniques," distributed in 2019, addresses the basic part of exact deals estimating in trade operations. Comparative ponders uncover that optimized ML models, such as Irregular Timberlands, can accomplish tall prescient exactness, beating conventional strategies like LR. For occasion, an optimized Arbitrary Timberland demonstrated accomplished an R-squared esteem of 0.945, considerably higher than conventional LR models.

- **(Comparison Study: Product Demand Forecasting with Machine Learning for Shop)_2_Reem :-**

  A consider titled "Comparison Study: Product Demand Forecasting with Machine Learning for Shop" assessed different calculations, counting K-Nearest Neighbor (KNN), Gaussian Credulous Bayes, and Choice Tree Classifier, to foresee item request. The discoveries shown classifier accomplished the Gaussian Naive Bayes classifier achieved the highest accuracy at 58.92%, outperforming KNN and Decision Tree Classifier, which had accuracies of 35.71% and 28.57%, respectively.

- **(Dynamic Model Selection Based on Demand Pattern Classification in Retail Sales Forecasting)_3_Reem :-**

  In the study "Dynamic Model Selection Based on Demand Pattern Classification in Retail Sales Forecasting," the authors evaluated the effectiveness of two dynamic weighting strategies—DWS-A and DWS-B—across four distinct demand patterns: smooth, intermittent, erratic, and lumpy. The performance of these strategies was compared against benchmark models, including Naïve, Comb S-H-D, and SCUM, using datasets from two retailers: Haolinju (an offline retailer) and JD (an online retailer).

  Smooth Demand Pattern:

  - *Haolinju Data (Forecast Horizon = 7 days):*
    - Naïve: sMAPE of 22.114%
    - Comb S-H-D: sMAPE of 18.749%
    - SCUM: sMAPE of 18.947%
    - DWS-A: sMAPE of 17.588%
    - DWS-B: sMAPE of 17.387%
  - DWS-B outperformed all other models, achieving a 5.23% improvement over SCUM.

  Intermittent Demand Pattern:

  - *Haolinju Data (Forecast Horizon = 7 days):*
    - Naïve: sMAPE of 82.682%
    - Comb S-H-D: sMAPE of 124.918%
    - SCUM: sMAPE of 126.654%
    - DWS-A: sMAPE of 75.052%

- DWS-A demonstrated an 11.1% improvement over the Naïve model.

Erratic Demand Pattern:

- *Haolinju Data (Forecast Horizon = 1 day):*
  - Naïve: sMAPE of 32.220%
  - Comb S-H-D: sMAPE of 31.539%
  - SCUM: sMAPE of 29.752%
  - DWS-B: sMAPE of 28.731%
- DWS-B achieved an 8.5% improvement over the Naïve model.

Lumpy Demand Pattern:

- *JD Data (Forecast Horizon = 7 days):*
  - Naïve: MASE of 75.817%
  - Comb S-H-D: MASE of 74.057%
  - SCUM: MASE of 74.092%
  - DWS-A: MASE of 64.560%
- DWS-A showed a 14.7% improvement over the Naïve model.

- **(Sales Demand Forecasting Using LSTM Network)_4_Reem:-**
  While thinking about "Sales Demand Forecasting Using LSTM Network," the creators executed a Long Short-Term Memory (LSTM) arrangement to anticipate future item requests. They employed a 60–40 data split, using 60% for training and 40% for testing. It is shown that the LSTM achieved a forecasting accuracy of 96.77%, beating other machine learning models assessed within the consideration.

- **Demand Forecasting Using Random Forest and Artificial Neural Network for Supply Chain Management" utilizes two models:(search_paper1_Tasneem):**
    - Random Forest (RF) && Keras Neural Network.

    The conclusion is :
        - Random Forest performed better than the Keras Neural Network in terms of predictive accuracy for demand forecasting in supply chain management.
        - The evaluation metrics used were R^2 score, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

    Random Forest Model:

    - R^2 Score: 0.83088
    - MSE Score: 0.00118
    - MAE Score: 0.02234

    Keras Neural Network Model:

    - R^2 Score: 0.81503
    - MSE Score: 0.00129
    - MAE Score: 0.02357
- **Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology.(search_paper2_Tasneem):**
    - used LSTM models to improve demand forecasting by capturing cross-product correlations in E-commerce data.
    - The models, trained across multiple related time series, achieved a Mean mMAPE between 0.424 and 0.455 on Walmart.com datasets, outperforming traditional forecasting techniques.

- The study concluded that LSTM models effectively enhanced prediction accuracy by leveraging time series grouping strategies.

- **An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise.(search_paper3_Tasneem):**
  - introduces a C-A-XGBoost model that combines clustering, ARIMA, and XGBoost to enhance sales forecasting.
  - The model consists of C-XGBoost for clustering-based forecasting, A-XGBoost for trend and residual modeling, and a weighted combination of both.
  - Performance evaluation showed that C-A-XGBoost outperformed ARIMA, XGBoost, and other models with the best accuracy.
  - The model achieved a Mean Error (ME) of 0.288, MSE of 10.769, RMSE of 3.282, and MAE of 2.515.
- **Procedia Computer Science 200 (2022) 993–1003" investigates the effectiveness of different models in sales forecasting, focusing on SARIMA, SARIMAX, and LSTM models. (search_paper4_Tasneem):**

  Models Implemented:
  - SARIMA , SARIMAX and LSTM .

  Results and Final Conclusion:
  - Salad: LSTM MAPE: 14, RMSE: 949; SARIMA MAPE: 13, RMSE: 1009.
  - Tomato: LSTM MAPE: 44, RMSE: 1170; SARIMA MAPE: 48, RMSE: 1163.
  - Potato: LSTM MAPE: 15, RMSE: 1184; SARIMA MAPE: 25, RMSE: 1626.
  - Cucumber: LSTM MAPE: 35, RMSE: 4502; SARIMA MAPE: 16, RMSE: 2854.

- For salad and tomato, SARIMAX improved prediction accuracy by reducing MAPE by 53% and 54%, respectively.
- Potato results showed a moderate improvement of 28%.
- Cucumber results improved by 12.5%, likely due to fewer promotional impacts.

- **(Effective Demand Forecasting Model Using Business intelligence with Machine Learning)_1_Mariam :-**

  According to the provided paper that was published in 2020 , It discussed the revenue losses problem that faced telecommunications companies due to customer churns then showed the effect on the machine learning models predictions in helping them to solve the problem. Also showed some machine learning models that give a good accuracy as the logistic regression and the random forest ensemble but it lacks interpretability. Due to the black box models as it gives the accuracy but does not know the main reason for this problem.Then it starts talking about the importance of hybrid models as “SHAPE” and XGBoost. Also it shows that these models give a high explainability and interpretability for the problem and also give an accuracy greater than the traditional models in the telecommunication industry. As the accuracy increased by 7% to reach 92% compared to the prior work.

- **(Explaining Machine Learning models in Sales Predictions)_2_Mariam :-**

  The problem of this paper is the same as last one but it focuses on the high dimensions data and how the models can deal with it. First in the data mining section it shows that the algorithm used for feature selection is a genetic algorithm that helps in identifying the top 10 important features. Then during the implementation of the used models they started to use

the decision tree model to get high interpretation and random forest ensemble learning for high accuracy. Last result of this paper it reached 88.5% accuracy and also it reduced the 15% of false positives compared with traditional work.

- **(Decision support system with ANN for demand forecasting)_3_Mariam :-**

  The problem discussed in this paper is the limited accuracy for the single classifier which is the decision tree, also the manual feature selection and the consumed time for this process. It starts to show the importance of hybrid approach and how it can provide a lot of advantages. It used the information gain and correlation-based filtering for the feature selection stage. Also use the ensemble model approche by combining Random Forest, AdaBoost, and Bagging to maximize accuracy. The result was that accuracy reached 82% as greater than the accuracy of a single classifier by 8%. Also the precision increased by 20%.

- **(Comparison of stochastic and machine learning methods for forecasting)_4_Mariam :-**

  Also this paper shows the problem of manual feature selection and the singler classifier limited accuracy .Then start to provide a hybrid model that can provide a comprehensive solution. It used the automated feature selection as it is a combination of information gain and correlation filtering. Also ensemble learning approaches by combining ANFIS with Bagging to enhance the performance. The results show that those techniques are more beneficial as the accuracy increased by 8% compared to singular classifier ro reached 0.948 AUC.