Project Synopsis

for

"SALES PREDICTION: AN INTEGRATED APPROACH USING MACHINE LEARNING"

A PROJECT SYNOPSIS

by

Japneet Singh
Aadika Bhatia
Lakshya Mishra
Manraj Singh

Introduction

Sales prediction is a critical task for businesses aiming to make informed decisions, optimize resources, and maximize profitability. With the advent of machine learning techniques, organizations can leverage historical sales data and other relevant factors to develop accurate forecasting models. This study presents an integrated approach to sales prediction, combining various machine learning methodologies to enhance prediction accuracy and reliability.

The objective of this study is to develop a robust model that can effectively forecast future sales based on historical data and relevant features. By utilizing machine learning algorithms, feature engineering, ensemble methods, and hyper parameter tuning, the integrated approach aims to provide organizations with valuable insights into future sales trends and patterns.

The study follows a step-by-step process, starting with data collection from the organization's historical sales records. This includes gathering information on sales periods, product attributes, customer demographics, pricing, and marketing efforts. Data preprocessing techniques are then employed to clean the data, handle missing values, and resolve inconsistencies.

Feature engineering plays a vital role in improving the predictive power of the data. Through the creation of new features or transformation of existing ones, the model can capture important temporal, lag, and domain-specific characteristics. These engineered features aim to capture underlying patterns and relationships that impact sales.

The selection of an appropriate machine learning algorithm is crucial to the success of the sales prediction model. Various algorithms such as Moving Average, linear regression, decision trees, random forests, gradient boosting, and neural networks can be considered based on the specific problem and dataset characteristics. The chosen algorithm is trained using the historical data, and ensemble methods are implemented to further enhance prediction accuracy.

Feature selection techniques help identify the most influential features for sales prediction, eliminating irrelevant or redundant variables. This step reduces complexity and improves the interpretability of the model.

The performance of the trained model is evaluated using standard evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared. Hyperparameter tuning techniques are applied to optimize the model's parameters and improve its performance.

Once the model demonstrates satisfactory performance, it is deployed in a production environment. Regular monitoring and iterative improvement ensure that the model remains accurate and up-to-

date with new data.

The findings of this study provide organizations with a reliable sales prediction model that aids in decision-making processes, resource allocation, and strategic planning. By harnessing the power of machine learning, businesses can optimize their operations and stay ahead in a competitive market landscape.

Objective

- **To predict the future sales** and use it as the basis of planning time and resources.
- **To maintain inventory** – An accurate sales forecasting helps in estimating the amount of materials required for future goals. It helps in keeping the inventory up for peak periods.
- Its helps in identifying Clients and Timelines:-In addition to what is selling, an accurate sales forecast can identify who is buying the most and when they are buying. If our sales are seasonal, we know which months are slow and can be used for prospecting.

Scope

Sales forecasting utilizes past figures to foresee present moment or long-haul execution. It's a questionable movement, in light of the way that such countless different factors can impact future deals: financial downturns, specialist turnover, changing examples and structures, extended test, creator surveys and various parts. Regardless, there are a couple of standard methods that can convey dependably exact deals gauges from year to year.

Without deals conjectures, it's uncommonly difficult for you to control the association the right way. You wouldn't understand that the spring is reliably the slowest season, so you'd put a great deal in stock that would just sit on the racks. You wouldn't concentrate on industry agents who envision an incredible improvement in event deals, and you'd lose potential customers to the test, which duplicated its get-away deals control and displaying endeavors. This exploration focuses:

- How significant is sales determining to the financial related arranging and the executives of a business?
- Techniques and innovations that guarantee the most precise and reliable forecasts?
- To structure a model to foresee the offers of thing through Machine Learning".
- To just limit the estimate blunder.

- In this we have old sales information of thing of any store and from that business information we need to foresee the future clearance of thing.

Literature Review

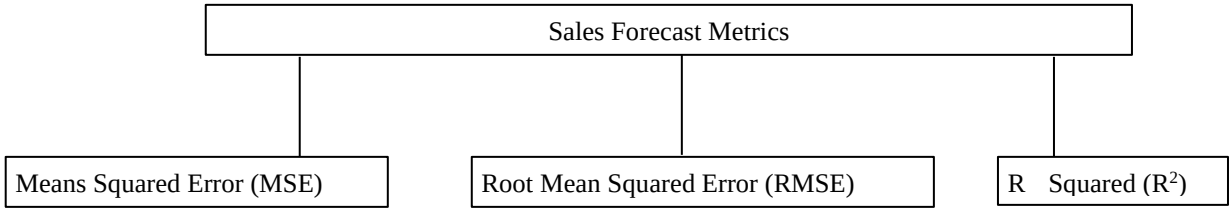
Sales forecasting and the approaches used, additionally giving a brief overview of DL models and metrics commonly used. According to Mentzer & Moon [2], a sales forecast is a "projection into the future of expected demand, given a stated set of environmental conditions". In some of the earlier works, instead of "sales forecast", the terms "sales prediction" or "demand forecast" have been used as synonyms of "sales forecasting". The general approach of using time series historical data for estimating the future value of sales is the common factor in the reviewed literature, therefore, for this study, "sales forecasting" as an umbrella term will be applied. For addressing forecasting problems, different models and methods have been used. Two of the classical forecasting methods are Auto-Regressive Integrated Moving Average (ARIMA), Seasonal Auto-Regressive Integrated Moving Average (SARIMA), where exponential smoothing performs statistical time series analysis. These are often used for market-level sales forecasts [3], [4].

Proposed Method

Each machine learning model takes care of an issue with an alternate target, utilizing a perfect dataset and consequently, it is imperative to comprehend the setting before selecting a metric. Ordinarily, the results to the accompanying inquiry can help us pick the suitable metric:

- Type of task: Regression?
- Business objective?
- What is the dissemination of the objective variable?

Relapse Metrics



Tools to be Used

HARDWARE:

Processor	:	I7
Memory	:	8GB RAM or above
Printer	:	Laser Printer
Pen Drive	:	5 GB

SOFTWARE:

Operating System	:	Windows 10.
Font-End Too l	:	Python
Back-End	:	Visual Studio Code

Python: Python is a popular programming language known for its flexibility, extensive libraries, and data analysis capabilities. It provides a wide range of libraries for regression analysis, such as Statsmodels and scikit-learn, which are suitable for implementing the study's methodology.

Pandas: Pandas is a powerful library in Python for data manipulation and analysis. It provides efficient data structures and functions for handling structured data, making it suitable for preprocessing and manipulating the sales data before applying regression analysis.

NumPy: NumPy is a fundamental library for scientific computing in Python. It provides powerful mathematical functions and an efficient array object that enables numerical operations. It is often used alongside Pandas for handling numerical data in regression analysis.

Statsmodels: Statsmodels is a Python library specifically designed for statistical modeling and analysis. It provides a wide range of statistical models, including regression models, and offers functions for estimation, hypothesis testing, and model diagnostics. It can be used for implementing and analyzing various regression techniques.

Scikit-learn: scikit-learn is a popular machine learning library in Python that offers a wide range of algorithms for regression, classification, clustering, and more. It provides an easy-to-use interface for implementing regression models, including linear regression, polynomial regression, and other

regression techniques. It also offers tools for evaluation and model selection.

Matplotlib and Seaborn: Matplotlib and Seaborn are Python libraries used for data visualization. They provide functions for creating various types of plots and graphs, allowing you to visualize the relationships between variables, analyze model performance, and present the results of the regression analysis.

Visual Studio Code: Visual Studio Code provides a user-friendly and customizable coding environment that is well-suited for Python development and data analysis tasks. By using VS Code, users can efficiently implement the proposed study on sales prediction, benefit from its features, and leverage its ecosystem of extensions to enhance our coding experience.

Result and Analysis

Principles:

1. Understand the problem before you begin to create the analysis model.
2. Develop prototypes that enable a user to understand how human machine interaction will occur.
3. Record the origin of and the reason for every requirement.
4. Use multiple views of requirements like building data, function and behavioral models.
5. Work to eliminate ambiguity.

A Complete Structure:

The limited time and resources have restricted us to incorporate, in this project, only the main activities that are performed in news sites, but utmost care has been taken to make the system efficient and user friendly.

For the optimum use of practical time it is necessary that every session is planned. Planning of this project will include the following things:

- Topic Understanding.
- Modular Break – Up of the System
- Processor Logic for Each Module.
- Database Requirements.

Topic Understanding:

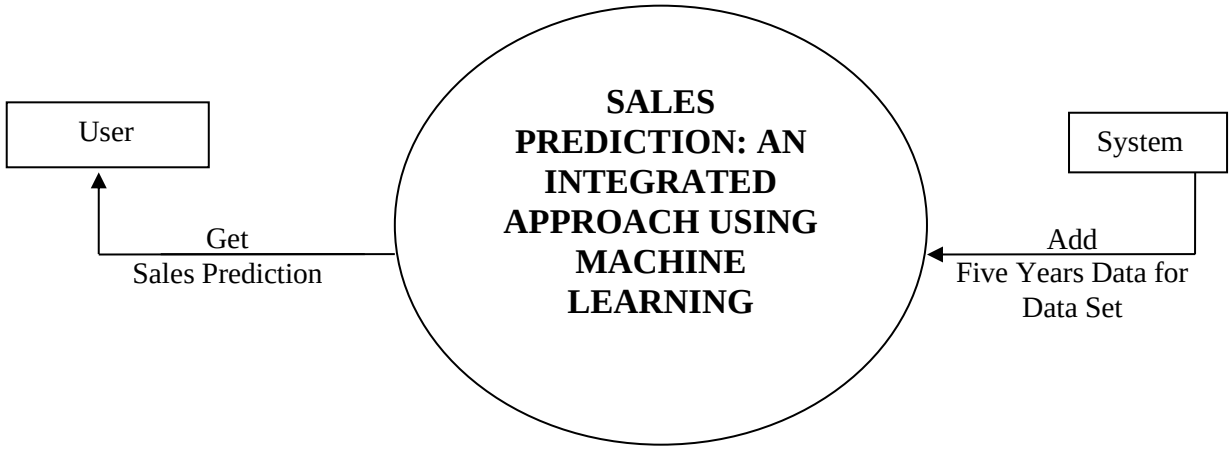
It is vital that the field of application as introduced in the project may be totally a new field. So as soon as the project was allocated to me, I carefully went through the project to identify the requirements of the project.

Modular Break –Up of the System:

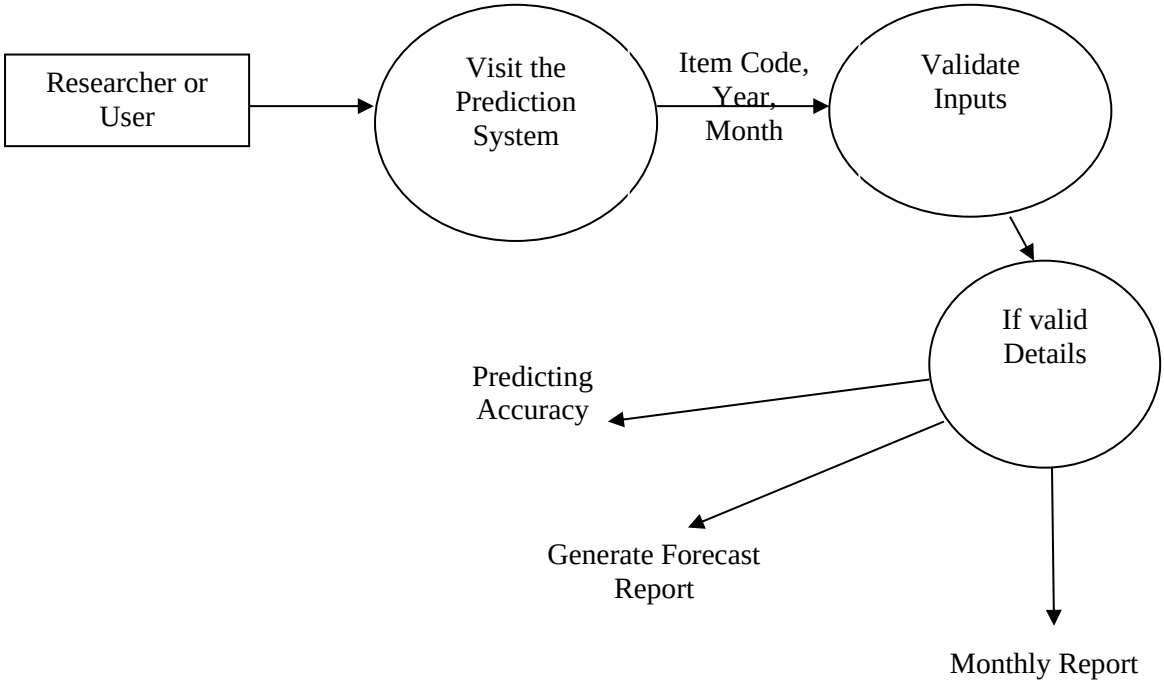
- Identify The Various Modules In The System.
- List Them In The Right Hierarchy.
- Identify Their Priority Of Development
- Description Of The Modules:

DATA FLOW DIAGRAM

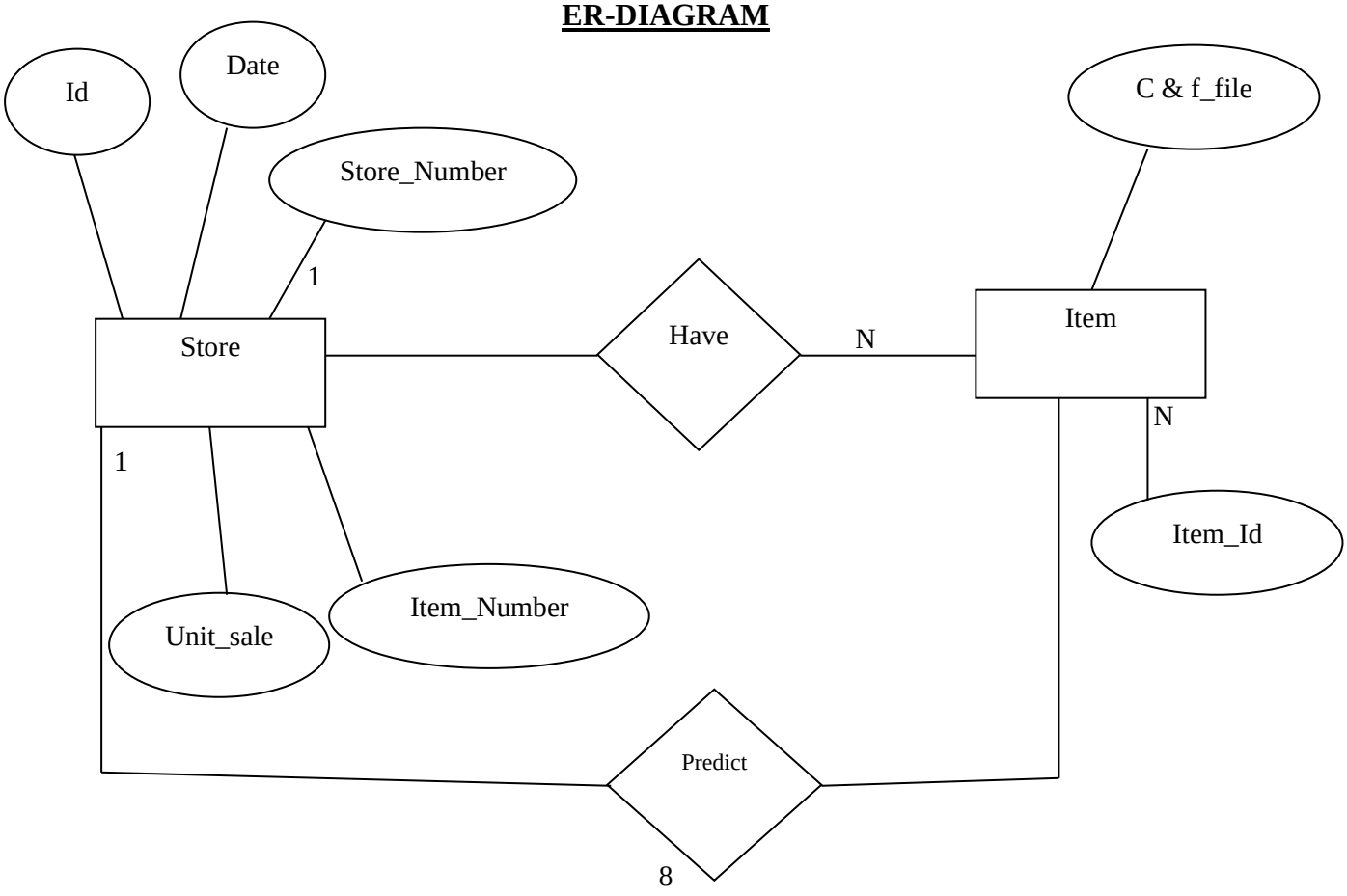
Context Level DFD



IST LEVEL DFD



ER-DIAGRAM



DATA TABLE

Dataset containing trained and test data is available on kaggle.com at https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data/.

Data is stored in the form of CSV files. These files contain comma separated values. A CSV file is used to store tabular data in the form of text file. Many online services allow its users to export tabular data from the website into a CSV file. We can open CSV file into Excel, and nearly all databases have a tool to allow import from CSV file. The standard format is defined by rows and columns data. Moreover, each row is terminated by a newline to begin the next row. Also within the row, each column is separated by a comma.

Train.csv file contains Ten lakh forty eight thousand records for twenty three stores. Each store is selling twenty items. The unit sale of each item in each store is give date wise.

ID	Date	Store_number	Item_number	Unit_Sales
0	1/1/2013	25	103665	7
1	1/1/2013	25	105574	1
2	1/1/2013	25	105574	2

Form train.csv, ten files are extracted having data for ten items.

S.No	File name	Items.No
1	Data2.cvs	956014
2	Data3.cvs	956013
3	Data4.cvs	956012
4	Data5.cvs	956011
5	Data6.cvs	119629
6	Data7.cvs	123601
7	Data8.cvs	153267
8	Data9.cvs	122725
9	Data10.cvs	153229
10	Data11.cvs	119624

Conclusion and Future Work

Most business specialists and data scientist concur that that sales forecasting should be a joint exertion. All around the best people to perform such activities are those most solidly included with the association's business works out.Contribution fuses direct relationship with customers, yet additionally an attention to economic situations. Counting key staff individuals from generation, stock administration and advertising advances a soul of cooperation and improves your capacity to make projections.

The expenses of data storage and storage security has made it feasible for little and medium sized organizations to deal with anticipating inside. Regardless, there are also various associations that you can contract with to support you. Discover them by securing referrals from your companions or checking trade disseminations. This study can help

- To minimize changeability
- Forecast results are improved
- Frameworks and methodologies are inter related
- Improves client administration
- Decreases lead time
- Better information of customers
- Better control of inventory
- Enables firms to respond all the more rapidly to changing economic situations.

Python language is used to create simulator to compare the results and to check the accuracy of the prediction system. **As Regression Metrics is more efficient to use. T**he Mean Squared Error (MSE) , R Squared (R^2) and Root Mean Squared Error (RMSE) are calculated**. Auto Regression Moving Average and Exponential smoothing algorithms in an Hybrid approach are giving the good results.**

References

1. M Giering, Retail sales prediction and item recommendations using customer demographics at store level-ACM SIGKDD Explorations Newsletter, 2021 - dl.acm.org.
2. Yi Yangl ; RongFulil ; Chang Huiyou ; Xiao Zhijiaol“SVR mathematical model and methods for sale prediction” Journal of Systems Engineering and Electronics Volume 18,pp 769-773,2019
3. Bohdan M. Pavlyshenko ,"Machine-Learning Models for Sales Time Series Forecasting",MDPI, January 2019
4. Jamal Fattah ,LatifaEzzine , ZinebAman, "Forecasting of demand using ARIMA model", International Journal of Engineering Management, 7 June 2018.
5. Maobin Li, Shouwen Ji and Gang Liu," Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA", Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model", Mathematical Problems in Engineering, Volume 2018.
6. Samaneh Beheshti-Kashi, "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering, Oct 2014.
7. Samaneh Beheshti-Kashi, "A survey on retail sales forecasting and prediction in fashion markets "Systems Science & Control Engineering An Open Access Journal 3(1):154-161 · January 2015
8. https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data
9. Ilham Slimani ; Ilhame El Farissi,"Artificial neural networks for demand forecasting: Application using Moroccan supermarket data", International Conference on Intelligent Systems Design and Applications (ISDA),2015.
10. **Ankur Pandey, Arun Chaubey ,Sanchit Garg, Shahid Siddiqui, Sharath Srinivas."Forecasting Demand for Perishable Items", 2012 (Nov)**
11. Xiao Fang Du, Stephen C.H. Leung ,Jin Long Zhang &K.K. Lai, "Demand forecasting of perishable farm products using support vector machine", Pages 556-567 | Received 08 Apr 2010, Accepted 06 Aug 2011, Published online: 10 Oct 2011
12. Rokach, L. Ensemble methods for classifiers. Data Mining and Knowledge Discovery Handbook; Springer: Cham, Switzerland, 2005; pp. 957–980.
13. Armstrong, J.S. Combining forecasts: The end of the beginning or the beginning of the end? Int. J. Forecast. 1989, 5, 585–588.

14. Papa charalampous, G.; Tyralis, H.; Koutsoyiannis, D. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water Resour. Manag. 2018, 32, 5207–5239

15. Taieb, S.B.; Bontempi, G.; Atiya, A.F.; Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst. Appl. 2012, 39, 7067–7083