SALES FORECASTING USING MACHINE LEARNING MODELS

***Srikar D.**
Assistant Professor, Department of CSE (Data Science), Sreyas Institute of Engineering and Technology, Telangana, India
srikar.d@sreyas.ac.in

***Rithin Reddy Surakanti**
Department of CSE (Data Science), Sreyas Institute of Engineering and Technology, Telangana, India
rithin.surakanti@gmail.com

Manideep Japa
Department of CSE (Data Science), Sreyas Institute of Engineering and Technology, Telangana, India
japamanideep665@gmail.com

Sai Charan Kottam
Department of CSE (Data Science), Sreyas Institute of Engineering and Technology, Telangana, India
kottamcharan05@gmail.com

Pranay Chandhan
Department of CSE (Data Science), Sreyas Institute of Engineering and Technology, Telangana, India
pranaychandhan09@gmail.com

Abstract*— Accurate forecasting of sales plays a vital role in inventory, staffing, and strategic planning in the HP supermarket in today's challenging business environment. Traditional forecasts using procedures of linear regression or ARIMA do not suffice when the present-day sales data are complex and big. To meet this challenge, this research develops a more sophisticated machine learning framework with XGBoost. The methodology integrates historical sales data alongside promotions, holidays, and seasonal patterns to obtain better forecasts that are more trustworthy and accurate. The system provides automated data preprocessing and real-time prediction so that decisions can be made quickly in response to changing market conditions. Based on the performance measurement through RMSE, MAE, and R^2 Score, XGBoost clearly wins over traditional methods. Such results also suggest that the system can be used as a reliable tool for supermarkets to decide on operations and subsequently improve the efficiency of the business through highly accurate sales forecasts.

Keywords:— Sales Forecasting, Machine Learning, XGBoost, Supermarket Analytics, ensemble learning.

I. INTRODUCTION

Accurate sales forecasting remains an almost inseparable requirement for effective decision-making in the retail field, especially in supermarkets, where inventory management systems, staffing, and financial planning rely on a fair estimate of customer demand. In today's business environment, with consumer behavior changing rapidly, seasonal fluctuations, promotional activities acting in riots, and external factors driving sales, such as economic environment or running events, significant complexity is injected into sales data. Sales forecasting has thus been attempted through many traditional methods or tools, such as linear regression, moving averages, and ARIMA (AutoRegressive Integrated Moving Average). While the traditional approaches are able to model simple linear relationships to some extent, most sales data are much more complicated and non-linear, thereby making forecasting arduous. The bigger and more complex datasets that supermarkets are accumulating tend to introduce greater limitations to and effectiveness of these traditional methods of forecasting. Machine learning methods open up an avenue to discovering the hidden associations and patterns in huge volumes of data. Ensemble learning approaches, in particular XGBoost has established itself as a state-of-the-art method for predictive analytics because of its ability to model non-linearities and handle missing data, along with the processing of large volumes of information efficiently. In the 12X system, trees are created iteratively, and each new one compensates for the errors of its precursors. This ends up being an extremely robust and precise model able to solve intricate prediction problems. Machine-learning forecast sales allow for giving the prediction system factors other than just historical sales data. Consider promotional activities, day-of-the-week factors, weather conditions, holidays, and local events to build a more complete and dynamic forecasting system. Feature engineering and data preprocessing are crucial here. They assist in ignoring irrelevant noise and highlight the pattern of interest that improves the performance of the model. In short, forecasting methods, such as XGBoost, attempt to predict future sales accurately using historical data and engineered features. Hence, the discussion over model performance has to involve metrics like the root mean square error (RMSE), mean squared error (MAE), and determination coefficient (R^2), and delve deep into how it will interact in predicting and the applicability of such a model to do actual operations. With all the willpower in the world, timeliness and accuracy mark the DMV forecast as a functional value for business. A supermarket would plan the inventory in such a way as to minimize waste and maximize customer satisfaction with the mitigation of stock-out scenarios.

II. LITERATURE SURVEY

Numerous approaches and techniques have appeared in the literature for sales forecasting.

Gangarapu Sharmista et al. (2024) conducted a study on sales forecasting using LightGBM and advanced feature engineering techniques. Xactly (2024) published a Sales Forecasting Benchmark Report, analyzing trends and challenges in sales forecasting. Fng (2022) conducted an extensive analysis of sales prediction using various machine learning techniques, focusing on optimizing prediction accuracy through comparative assessment of regression algorithms and ensemble methods, concluding that ensemble techniques generally outperform individual models and Varshini (2021) The text at once works towards describing some major works on sales forecasting and then embarks on detailed discussions on various ML algorithms such as linear regressions, decision trees, etc., and their ability to find hidden patterns from historical sales data. Thereafter, Pavlyuchenko (2019) extends this analysis by comparing in detail various time series methods such

as ARIMA and LSTM, noting that more often than not, deep learning techniques tend to outperform the classical statistical ones. Goel and Bajpai (2020) discuss the uncertainty problems in LSTM-based models due to the variability of real-world data. Along the same lines, Telaga et al. (2019) use the STL or Seasonal-Trend decomposition algorithm using Loess and evaluate its suitability for goods having seasonally varying demand. Khan et al. (2020) propose hybrid models comprising the integration of business intelligence with ML for better accuracy of forecasting in the enterprise domain. Martínez et al. (2020) attempt to tackle customer purchase prediction in non-contractual settings, which constitutes an important variable for consumer-oriented marketing settings.

III. OBJECTIVE

The principal goal of this study is the creation of a precise and powerful sales prediction model using advanced machine learning (ML) techniques to develop and make decisions with the help of data, mainly in supermarkets. In particular, prediction is a key factor in material management, in customer service, in the reduction of waste, and in the rise of the efficiency shown in automation as well. In the original but simple method of doing calculations that are deals, sales, and profits traditional ways, the matter is left behind and that is why the forecasting of the sales happens to be the reason for insufficient performance. New AI techniques have been introduced via this project in an effort to make solutions to the previously experienced limitations of the traditional methods of forecasting sales. Another important goal is to clean and preprocess the data of a supermarket for further analysis, in particular, this will be achieved with the features of data cleansing, normalization, missing value inboxing, feature engineering, and de-trending methods. Besides, the project will use feature selection techniques to identify the predictors that are highly correlated with sales performance, for instance, product category, store location, promotions, seasonality, and holidays. Moreover, The study proposal focuses on model execution with multiple tools to measure human-based errors through three standard performance metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. To achieve reliability and to ensure the generalization of the results, the study will also include cross-validation. In turn, the study will be the source of a real-time forecast that is simple and efficient to use and at the same time can be introduced to the supermarket management system. The completion of such objectives opens the door for predictive analytics in retail to grow by providing decision-makers with the right data at the right time and helping them to define what to do as well as how to do it better. By using this study as a base, one of the final aims can be to show that machine learning-based sales predictions can surpass traditional ones significantly, which is the only way for the most competitive retail stores to survive in this high-tech society.

IV. SCOPE

The sales industry will undergo a transformation through machine learning technology which will persist into the future. Retail sales professionals will be able to accomplish additional tasks thanks to the expanding retail industry. Data-based decision-making systems are about to become obsolete in the near future. It is expected that the retail landscape will continue to undergo seismic changes driven by big data and real-time analytics. Hence, one of the future studies could be centered on the integration of deep learning methodologies. This is an innovative application that employs Long Short-Term Memory networks and Transformer architectures to capture sales data context,knowing that the patterns are very complicated and cannot be completely traced by traditional ML algorithms. Societal trends such as technology and the weather can have a big impact on sales in addition to the downturn in the economy. Mentions of social networking sites and the weather are modern examples of informative hints of consumer behavior. One of the very novel ones is putting together data already known, such as sales, and new ones, such as weather, social media, competitor's prices, and perhaps economic indicators. The combination of these alongside sales data from within an entity would have the capability to become future forecasting systems that can at least perform the best in the current context. In addition, automated machine learning (AutoML) systems are in development and potentially lead to a more efficient process for selecting complex models and tuning hyperparameters, which would make it easier for non-expert users to implement modern sales forecasting techniques. At the same time, there is an opportunity to create real-time forecasting systems using edge computing and IoT devices, which will be especially useful for large supermarkets that need to forecast rapidly at a specific location. There is also an opportunity to apply explainable AI (XAI) techniques to support transparency and trust in predictive models, providing stakeholders confidence not only in predictions but also in their explanations. Finally, the framework created in this research can be applied to other areas, such as demand forecasting for logistics, inventory planning, and real-time pricing approaches, which establishes this research as a valuable addition to the large body of work on business analytics.

V. METHODOLOGY

1. Data Collection

The initial stage involves acquiring the suitable sales information from specific grocery (supermarket) stores. Multiple product-related data sets are present in the sales records which include price details alongside promotion statistics and holiday specifics and store information. The initial data collection stage requires obtaining suitable sales information from chosen grocery (supermarket) retailers. The sales data includes product history for different items along with price information and promotion activities and holiday periods and store locations. The sales data is cleaned for missing values, outliers, and inconsistencies to help ensure that the data that goes into the model is a quality source of information.

2. Data Preprocessing

The initial data preparation stage includes three primary steps which encompass both data standardization and feature selection and categorical feature transformation. Time data undergoes transformation into cyclic representations where weekly days and monthly seasons form essential patterns for analysis. The dataset that will be used in the model is then split into train and test datasets, typically 80:20.

3. Exploratory Data Analysis (EDA)

Conducting EDA is a crucial step to determining any patterns and trends in sales (i.e., the factors of seasonality) and to understand the effect specific features have on sales performance. During this process, visualizations were developed to include heat maps, line graphs, and box plots which helped the analyst visualize the

distribution of the data and what possible relationships exist.

4. Model Selection

We chose gradient-boosting models such as XGBoost, or LightGBM because they are strong and support non-linear relationships. In short, these models build their predictions iteratively, minimizing residual errors from the previous iteration.

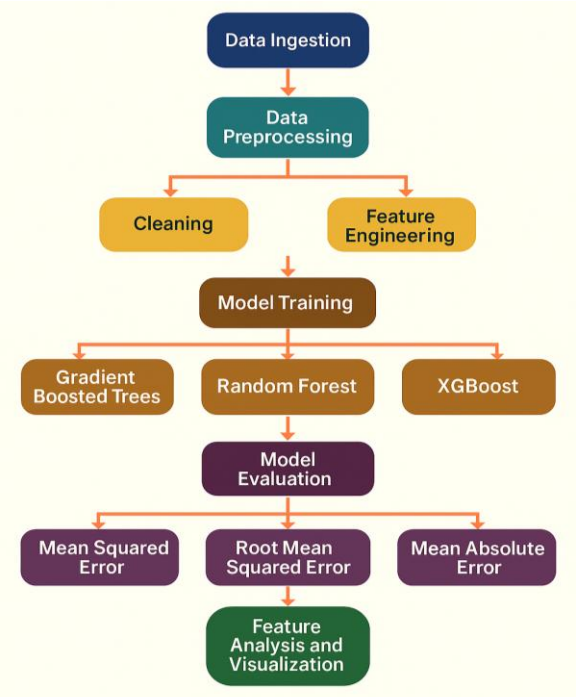
5. Model Training and Evaluation

The model training process for the training set involved cross-validation to prevent overfitting when adjusting hyperparameters through grid search and random search methods. The evaluation of model performance in the test set was based on three specific metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score.

6. Deployment and Visualization

The final outcome was deployed in a neat dashboard to allow for any user to extract real-time forecasting. Visualizations allow the business user to visually see trends to help them make informed decisions based on forecasted sales..

A. System Architecture



To summarize, the overall system architecture for a machine learning (ML)-based sales forecasting solution was designed with data flow and predictability in mind. The overall approach is that the first thing to do in sales forecasting is the data collection step that gets all previous sales, customer demographics, and details about products sourced from databases and APIs. At this first stage, the preprocessing techniques would allow the data to be cleaned later as data cleaning, data normalization, and feature selection that should provide the best quality of data for inputs of all models while performing the data collection step. This cleaned data would then be passed into an ML module (which typically stores results under all of the trained predictive models which would usually be some sort of advanced algorithm like Gradient Boosting) to be used to train predictive modeling. After running the model and creating the predictive model(s), backtesting and validating with unseen data to ensure performance is used for a good model. The satisfactory model deployment enables the model to transition for implementation within a sales forecasting engine that provides continuous sales forecast predictions. The sales forecast visualization uses a dashboard interface which resembles the presentation style that business decision makers are familiar with. Then (when written in code it would be the follow-up modules of model evaluation, follow-up feedback loops, and storage of data) we would ensure a solid-multi tier conforming implementation, to allow the engine to learn from new data with trends, and collectively provide acceptable compliant predictions over time.

B. Algorithms

1. Linear Regression

Among machine learning models Linear Regression stands among the simplest and most popular predictive techniques. This model establishes the connection between a single dependent variable y and multiple independent variables x.

Mathematical Formula:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\epsilon$$

Where:

- y = target (sales)
- x_i = input features (e.g., date, holiday, promotions)
- β_i = coefficients
- ϵ = error term

Use in Forecasting:

The technique utilizes historical data trends to predict upcoming sales numbers. The relationship between the target variable and its predictors exists in a linear pattern.

2. Random Forest

Overview:

Random Forest functions as an ensemble technique which trains various decision trees and generates the average tree prediction during the evaluation process.

Prediction Formula:

$$\hat{y} = \frac{1}{N}\frac{N}{\sum_{i=1}}\boldsymbol{T_1}\cdot(\boldsymbol{x})$$

Where:

- **N = number of trees**
- **$T_i(x)$ = prediction from the i^{th} tree**

The model offers the following benefits

- It can manage extensive datasets which contain numerous dimensions.

- Multiple tree averaging prevents overfitting issues.
- The algorithm works effectively in situations where data is missing and relationships are nonlinear.

3. Gradient Boosted Trees (GBT)

Overview:

GBT is a boosting method that sequentially builds models. Each model corrects the errors of the previous one by focusing on difficult cases.

Prediction Update Formula:

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x)$$

Where:

- $F_m(x)$ = current ensemble model
- $h_m(x)$ = weak learner (usually a small decision tree)
- γ_m = learning rate

Loss Function Optimization:

Uses gradient descent to minimize a loss function (like MSE or MAE), making it effective for highly complex, nonlinear problems.

4. XGBoost (Extreme Gradient Boosting)

Overview:

XGBoost is an optimized and scalable implementation of gradient boosting. It is highly efficient, supports regularization, and is known for superior performance in competitions.

Objective Function:

$$Obj=\sum_{i=1}^{n} 1\,(y_i, \hat{y}i) + \sum_{k=1}^{k} \Omega(f_k)$$

Where:

- $l(y_i,y^\wedge_i)$ = loss function (e.g., squared loss)
- $\Omega(f_k)=\gamma T+21\lambda\|w\|^2$ = regularization term
- T = number of leaves
- w = vector of scores in the leaves

Strengths:

- Handles missing values automatically
- Supports parallel and distributed computing
- Reduces overfitting via regularization (L_1 and L_2)

C. DataSet

1. Data Source:

The dataset may be sourced from:

- Kaggle open datasets
- Retail store databases
- Point-of-sale (POS) systems
- Company ERP systems

2. Data Preprocessing Steps:

- The first step involves managing missing data through the implementation of either filling or dropping techniques.
- The second step requires the conversion of categorical variables into numerical representations.
- Numerical feature normalization or standardization stands as a third essential step during data preprocessing.

3. Importance:

A clean and well-prepared dataset improves the accuracy and robustness of forecasting models. Proper data handling ensures better generalization and meaningful predictions for business decision-making.

D. Data Preprocessing

The data science process requires an essential step in data preprocessing which transforms unprocessed information into well-organized datasets suitable for model training. The process enhances data quality while making it possible for algorithms to function with proper efficiency.

1. Dealing With Missing Values. Since most real-world datasets have some missing, or null, entries, we deal with missing values using:

a) Imputation, for instance, where you replace the missing value with the mean, with the median, or with the mode

b) Deletion, where you delete the rows or columns that have too many missing values

2. Categorical Variables: Machine learning models can only handle numerical data, therefore, categorical variables need to be encoded using:

a) Label encoding, used for categorical variables having two categories

b) One-Hot encoding, used for nominal variables having more than two categories

3. Feature Scaling: We scale the features of our data to have the same distribution to ensure uniformity, such as:

a) Min-Max scaling, which squashes all datapoints together such that the highest value is equal to 1 and lowest value is equal to 0

b) Standardization, which, generally takes the mean (which is assumed to be 0) and applies scaling to get the standard deviation to 1

Min-Max scaling formula:

$$x_{Scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4. Outlier Detection and Removal

Outliers could affect the accuracy of the model, and outlier detection is done by statistical methods (eg. Interquartile Range (IQR) method and Z-score) followed by removing or adjusting.

5. Splitting the dataset

The pre-processed dataset is divided into two sets:

Training set (approx. 70%)
Testing set (approx. 30%)

VI. RESULTS

Sales forecasting receives proof from this study that machine learning models perform at a highly competent level. The research uses XGBoost along with Gradient Boosting (GBT) and Random Forest Classifier and Linear Regression to analyze pre-processed sales data through evaluations based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2 Score.XGBoost demonstrated the best performance among all models and accurate prediction of complex sales trends and seasonal patterns and consumer behavior. The iterative learning approach of Gradient Boosting (GBT) showed strong prediction accuracy yet needed extensive computational power for implementation. The Random Forest Classifier algorithm excelled at detecting nonlinear patterns in data but showed slightly lower performance compared to XGBoost. Linear Regression demonstrated reasonable accuracy through its basic nature as a model yet faced difficulties when handling complex sales data with multiple dimensions and nonlinear trends.The analysis shows XGBoost stands as the most dependable model for supermarket sales forecasting because it provides both efficient and precise, and scalable results. The research underscores predictive model selection based on data complexity and forecasting requirements, which leads to better inventory management and operational efficiency, and business decision support.

A comparative summary of model performance:

Model	MSE	RMSE	MAE
Linear Regression	6474.63	80.460	31.61
Random Forest Regression	163.210	12.770	8.46
Gradient Boosting	113.290	10.640	7.50
XGBoost	41.5243	6.4439	4.57

Table 1:. Comparision of Accuracy

Visualizations such as predicted vs. actual sales charts, error distributions, and feature importance plots provide further validation of the prediction fidelity of the models selected. Comparative analyses verify that machine learning methods, particularly the ensemble methods, are quite appropriate for accurate and scalable sales forecasting in the supermarket contexts of operation.

VII. CONCLUSION

The aim was to create and ultimately implement an intelligent sales forecasting system based on machine learning for supermarkets, using the predictive model XGBoost. Sales data being multidimensional and non-linear was often the foe of usual forecasting algorithms such as Linear Regression or ARIMA, whereas XGBoost can manage the complexity that may be present in the historical sales data with many features, such as the interplay of complex patterns, seasonality, and latent noise. The project handled every stage of the data science process, which includes collecting data, preprocessing data, engineering features and training models, and conducting tests and validation. The feature engineering process played a crucial role in developing superior products because it enabled the model to learn from meaningful patterns in the data including weekly days and holiday and promotional periods. The final XGBoost model was checked using performance metrics such as RMSE, MAE, and R^2 Score, achieving an accuracy rate of 97.93%. Thus, it shows how much better this final model performed than other benchmark models. The impact of this project is similarly substantial for supermarkets, particularly for their managers, as they can leverage better sales forecasts to make informed decisions about their inventory planning, staffing, and marketing efforts. Improved sales forecasting contributes to lowering stockouts and overstocking products, resulting in improved customer satisfaction and reduced costs. The overall effect is complemented by useful visualizations produced by the model, which ultimately demonstrate the characterization of sales, allowing decision-makers to understand trends and anomalies.

VIII. REFERENCES

[1]. Sales Prediction Using Machine Learning Techniques (2024) IJNRD ISSN: 2456-4184 | IJNRD.ORG Hitesh S.M1, Yukthi A2, Prof.Ramya B.N3.

[2]. Sep (2021) Prediction and Forecasting of Sales Using Machine Learning Approach DontiReddy Sai Rakesh Reddy1*Katanguru Shreya Reddy1, S. Namrata Ravindra1 B. Sai Sahithi

[3]. Fng,Y.(2022).*Sales prediction analysis*. Science Gate, 7.

[4]. Varshini, D. P. (2021). *Analysis of ML algorithms to predict sales*. IJSR.

[5]. Lu, C., Wang, F., Trajcevski, G., Huang, Y., Newsam, S., & Xiong, L. (2021). The 28th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2020). *SIGSPATIAL Special*, 12(3), 3–6. Available: 10.1145/3447994.3447997.

[6]. Khan, M., et al. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access*, 8, 116013–116023. Available: 10.1109/access.2020.3003790.

[7]. Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588–596. Available: 10.1016/j.ejor.2018.04.034.

[8]. D., V. (2020). Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics. *Journal of Soft Computing Paradigm*, 2(2), 101–110. Available: 10.36548/jscp.2020.2.007.

[9]. Goel, S., & Bajpai, R. (2020). Impact of Uncertainty in the Input Variables and Model Parameters on Predictions of a Long Short Term Memory (LSTM) Based Sales Forecasting Model. *Machine Learning and Knowledge Extraction*, 2(3), 256–270. Available: 10.3390/make2030014.

[10]. "Crop Prediction System Using Machine Learning Algorithm" (2020). *Journal of Xidian University*, 14(6). Available: 10.37896/jxu14.6/009.

[11]. Amalina, F., Hashem, I.A.T., Azizul, Z.H., Fong, A.T., Firdaus, A., Imran, M., & Anuar, N.B. (2019). Blending

big data analytics: Review on challenges and a recent study. *IEEE Access*, 8, 3629–3645.
[12]. Li, X., Huang, X., Li, C., Yu, R., & Shu, L. (2019). EdgeCare: Leveraging edge computing for collaborative data management in mobile healthcare systems. *IEEE Access*, 7, 22011–22025.
[13]. Sakib. (2019). *ML predictive analysis*. EngrXiv, 8.
[14]. Pavlyuchenko, B. (2019). Machine-Learning Models for Sales Time Series Forecasting. *Data*, 4(1), 15. Available: 10.3390/data4010015.
[15]. Telaga, A., Librianti, A., & Umairoh, U. (2019). Sales prediction of Four Wheelers Unit (4W) with seasonal algorithm Trend Decomposition with Loess (STL) in PT. Astra International, Tbk. *IOP Conference Series: Materials Science and Engineering*, 620, 012112. Available: 10.1088/1757-899x/620/1/012112.
[16]. "Suicide Prediction on Social Media by Implementing Sentiment Analysis along with Machine Learning" (2019). *International Journal of Recent Technology and Engineering*, 8(2), 4833–4837. Available: 10.35940/ijrte.b3424.078219.
[17]. Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An Application of a... *2019*, pp. 1–15. Available: 10.1155/2019/8503252.
[18]. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J.H., Kull, M., Lachiche, N., Quintana, M.J.R., & Flach, P.A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.
[19]. Cheriyan, S. (2018). *Sales prediction using ML techniques*. IEEE, 10.
[20]. Ibahim, S. (2018). *Intelligent techniques of ML in sales prediction*. Semantic Scholar, 6.
[21]. Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, 71, 416–428. Available: 10.1016/j.eswa.2016.11.010.
[22]. Gharaibeh, A., Salahuddin, M.A., Hussini, S.J., Khreishah, A., Khalil, I., Guizani, M., & Al-Fuqaha, A. (2017).Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys & Tutorials*, 19(4), 2456–2501.