AI and Predictive Analysis: A Case Study of Customer and Transaction Data

Hari Prasad Bomma

Data Engineer, USA
haribomma2007@gmail.com

Abstract

Predictive analytics involves using data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on past data. It's about making informed guesses about what might happen next. The exploration of AI-based predictive analysis on customer and transaction data represents a significant opportunity for companies. This research paper explores the application of AI-based predictive analysis techniques to customer and transaction data, highlighting the potential benefits and challenges of this approach. In this paper we will see a case study that used AI predictive analysis to enhance the decision-making and improving customer experience.

Keywords: AI, Predictive analytics, Descriptive Analytics, Diagnostic Analytics, Prescriptive Analytics, Regression Models, Decision Trees, Ensemble Methods, Neural Networks, Support Vector , Machines (SVM), Bayesian Models, Clustering Models, Time Series Models, Association Rule Learning

Introduction:

Predictive analytics is a powerful approach that combines historical data, statistical models, and machine learning techniques to forecast future events. By analyzing patterns and trends from past data, predictive analytics helps organizations anticipate outcomes, identify potential risks, and seize opportunities. It's widely used across various industries to optimize marketing strategies, improve operational efficiency, and enhance decision-making processes. Essentially, it transforms raw data into actionable insights, enabling businesses to make more informed and strategic decisions about what may happen next.

In the era of digital transformation, companies are increasingly relying on data-driven insights to enhance customer experience and drive business growth. Artificial intelligence has emerged as a powerful tool for predictive analysis, enabling organizations to uncover valuable patterns and trends from vast customer and transaction datasets [1].

The rapid advancements in artificial intelligence and the availability of large datasets have revolutionized the field of business analytics. New technologies such as big data and AI have enabled companies to perform predictive, descriptive, and prescriptive analyses, providing them with a competitive edge. [2] As the paper notes, "These technologies represent a revolution in how companies can obtain a great quantity of valuable data and how they analyze them."[2] The adoption of AI-driven systems has had a significant impact on various business contexts, including customer interaction, sales platforms, and employee skill sets. [3]

Data Analytics: Data Analytics is the overall field that involves examining data sets to draw conclusions about the information they contain. Data analytics is the process of examining data sets to draw conclusions about the information they contain. It involves various techniques and tools to discover patterns, correlations, and trends within data. The goal is to transform raw data into meaningful insights that can support decision-making and strategic planning. It is widely used across various fields, including business, healthcare, finance, and marketing, to gain valuable insights and enhance performance.

 Descriptive Analytics: Focuses on summarizing historical data to understand what has happened in the past. Techniques include data aggregation, data mining, and reporting tools such as dashboards and visualizations. It helps organizations identify patterns and trends, providing a clear picture of past performance.

 Diagnostic Analytics: Explores data to determine the reasons behind past outcomes. It involves identifying root causes and understanding the relationships between variables. Techniques include drill-down, data discovery, data mining, and correlations. Diagnostic analytics answers questions like "why did this happen?" and helps in problem-solving and troubleshooting.

 Predictive Analytics: Uses historical data, statistical models, and machine learning techniques to forecast future outcomes. It involves the use of algorithms, predictive modeling, and data mining to identify patterns and make predictions. Common applications include demand forecasting, risk assessment, and customer behavior prediction.

 Prescriptive Analytics: Suggests actions to achieve desired outcomes or solve problems, often using predictive models. It combines data analysis, optimization algorithms, and simulation to recommend the best course of action. Prescriptive analytics helps organizations make informed decisions by providing actionable recommendations based on data-driven insights. Applications include supply chain optimization, resource allocation, and personalized marketing strategies.

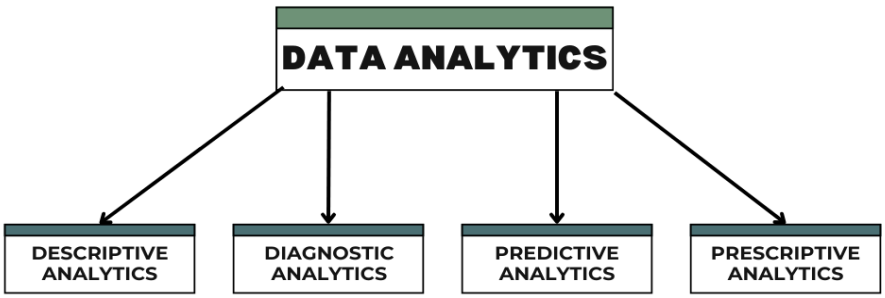

Figure 1: Data Analysis Techniques

Predictive Analytics: Models and Key Techniques:

Models: Models are the actual algorithms or systems built using one or more techniques to make specific predictions. Models are created by training on historical data and are used to forecast future outcomes. There are different models are in predictive analytics. These are selected based on the nature of the problem, the type of data available, and the desired outcome.

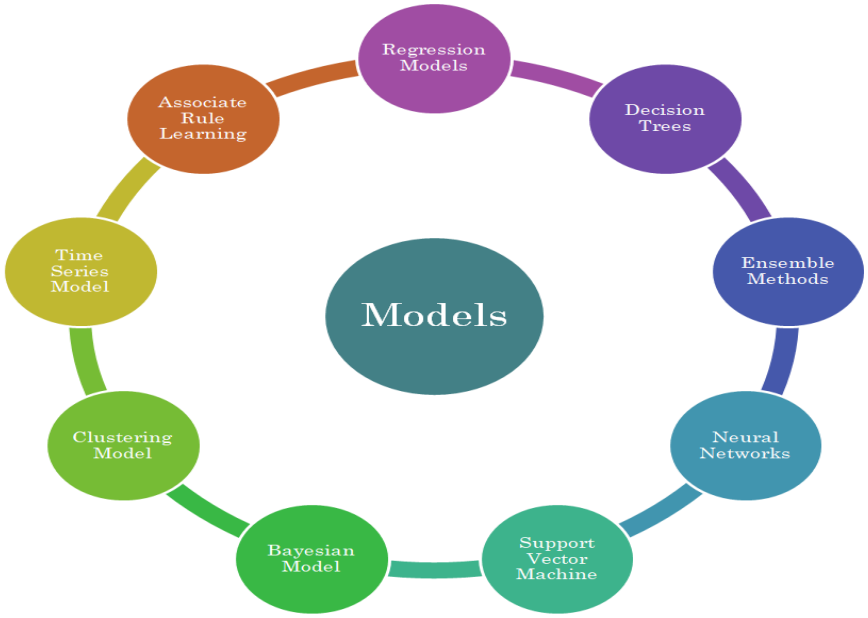

Figure 2: Predictive Analytics Models

- **Regression Models**: Used to predict continuous outcomes by establishing relationships between dependent and independent variables. Examples include linear and logistic regression.
- **Decision Trees**: Utilize a tree-like model of decisions and their possible consequences. They are used for both classification and regression tasks.
- **Ensemble Methods**: Combine multiple models to improve predictive performance. Examples include Random Forest and Gradient Boosting Machines (GBM).
- **Neural Networks**: Composed of interconnected nodes (neurons) that process data in layers. This is used for complex pattern recognition and prediction tasks.
- **Support Vector Machines (SVM)**: Identify the optimal hyperplane that separates data into different classes, used for both linear and non-linear classification tasks.
- **Bayesian Models**: Based on Bayes' theorem, these models calculate probabilities to make predictions. Examples include Naive Bayes and Bayesian Networks.
- **Clustering Model**: Groups data points based on feature similarity. Examples include K-Means Clustering and Hierarchical Clustering.
- **Time Series Models**: Analyze data points collected over time to forecast future values. Examples include ARIMA and Exponential Smoothing.
- **Association Rule Learning**: Identify relationships between variables in large datasets. An example is the Apriori Algorithm.

Techniques: These are the methods or approaches used to analyze data and make predictions. Techniques involve the specific statistical and computational methods applied during the analysis process.

- **Regression Analysis**: These are statistical methods that examine the relationship between a dependent variable and one or more independent variables. They help to understand how the typical value of the dependent variable changes when any one of the independent variables is varied.
- **Decision Trees**: This technique uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each internal node represents a test on

an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or decision.

- **Machine Learning**: These algorithms allow computers to learn from data and make predictions or decisions without being explicitly programmed to perform the task. Machine learning techniques include supervised learning (where the model is trained on labeled data), unsupervised learning (where the model finds hidden patterns in unlabeled data), and reinforcement learning (where the model learns by receiving rewards or penalties). Common applications include image recognition, natural language processing, and recommendation systems.
- **Time Series Analysis**: This technique involves analyzing sequences of data points collected or recorded at specific time intervals. It's used to identify patterns over time and forecast future values based on historical data. Examples include stock market analysis, weather forecasting, and demand prediction. Key methods in time series analysis include moving averages, exponential smoothing, and ARIMA models (AutoRegressive Integrated Moving Average).

Predictive Analysis on Customer and Transaction Data:

Customer and transaction data are the lifeblood of many businesses, containing a wealth of information about consumer behavior, preferences, and purchasing patterns. AI-powered predictive analysis techniques can be used to extract valuable insights from these datasets, enabling companies to improve forecasting, optimize operations, develop targeted marketing strategies, and enhance the user experience. However, unlocking the full potential of AI-based predictive analysis requires a robust digital foundation and the right data ecosystem. Companies must invest in building the necessary infrastructure, including data management systems, analytical tools, and skilled personnel, to effectively leverage AI in their decision-making processes.

Case Study: Enhancing Customer Insights through Transaction Data Analysis. Sample data and example are considered for this paper to show the results.

Background:

A mid-sized retail chain, sought to improve its customer relationship management (CRM) by using transaction data to gain deeper insights into customer behavior and preferences.

Objective:

To analyze the relationship between customer data and transaction data to enhance customer segmentation, predict customer lifetime value (CLTV), and identify potential churn risks.

1. Customer Lifetime Value (CLTV) Prediction: Analyzing the total amounts spent and the frequency of visits, we can predict the future value of each customer.

2. Churn Prediction: Using recent visit dates and transaction amounts, we can identify customers who are at risk of churning (i.e., stopping their transactions).

3. Purchase Pattern Analysis: We can determine purchasing patterns, such as preferred payment methods and discount types.

Data Collection:

Customer Data: Collected from CRM system, including customer ID, first name, last name, address, state, zip code, country, first visit date, recent visit date, status indicator, record start date, and record end date.

ID	F_Name	L_Name	Full_name	Add1	Add2	State	Zipcode	Country	First_Visit	Recent_Visit	Status_Ind	Record_SDate	Record_EDate
1	Alex	Johnson	Alex Johnson	123 Elm St	Apt 5A	TX	75001	USA	6/15/2021	11/30/2022	Active	6/1/2021	6/1/2025
2	Sarah	Davis	Sarah Davis	456 Maple Ave	Suite 12	CA	90210	USA	3/10/2020	12/1/2022	Active	3/1/2020	3/1/2024
3	Michael	Brown	Michael Brown	789 Oak St	Floor 2	NY	10001	USA	11/25/2019	10/20/2022	Inactive	11/1/2019	11/1/2023
4	Emily	Wilson	Emily Wilson	101 Pine St	Apt 6B	FL	33101	USA	7/22/2020	9/15/2022	Active	7/1/2020	7/1/2024
5	David	Lee	David Lee	234 Cedar Rd		IL	60601	USA	2/1/2021	8/30/2022	Inactive	1/15/2021	12/31/2023
6	Linda	Martinez	Linda Martinez	567 Birch Dr	Suite 3	OH	44101	USA	9/12/2020	7/20/2022	Active	9/1/2020	1/31/2024
7	Robert	White	Robert White	890 Spruce Ln	Apt 4C	GA	30301	USA	5/18/2019	6/10/2022	Active	5/1/2019	5/1/2024
8	Maria	Gonzalez	Maria Gonzalez	123 Aspen Ct	Floor 1	TX	75201	USA	8/8/2018	5/22/2022	Inactive	8/1/2018	8/1/2023
9	James	Clark	James Clark	456 Walnut St	Apt 7D	NY	10002	USA	1/5/2021	4/30/2022	Active	1/1/2021	1/1/2024
10	Anna	Roberts	Anna Roberts	789 Redwood Blvd	Suite 5	CA	90211	USA	4/15/2022	3/20/2022	Active	4/1/2022	4/1/2025

Table 1: Customer Table

Transaction Data: Collected from point-of-sale (POS) systems, including transaction ID, customer ID, net amount, payment type, total amount, discount amount, discount type, and invoice number.

Tran_ID	Cust_ID	Net_Amt	Pmt_Type	Total_Amt	Discount_Amt	Discount_Type	Invoice_Num
101	1	150	Credit	135	15	Percentage	INV001
102	2	200	Cash	190	10	Fixed	INV002
103	3	350	Debit	315	35	Percentage	INV003
104	4	120	Credit	108	12	Percentage	INV004
105	5	500	Cash	450	50	Percentage	INV005
106	6	250	Debit	225	25	Percentage	INV006
107	7	300	Credit	270	30	Percentage	INV007
108	8	400	Cash	380	20	Fixed	INV008
109	9	600	Debit	540	60	Percentage	INV009
110	10	700	Credit	630	70	Percentage	INV010

Table 2: Transaction Table

1. Customer Lifetime Value (CLTV) Prediction:

- **High-Value Customers**: Customers like Linda Martinez (ID: 6) and James Clark (ID: 9) consistently make higher-value purchases and are predicted to continue this trend, making them high-value customers.

CLTV Calculation: Use a simple formula to calculate CLTV:
CLTV= (Average Purchase Value×Purchase Frequency)×Customer Lifespan\text{CLTV} = (\text{Average Purchase Value} \times \text{Purchase Frequency}) \times \text{Customer Lifespan}

Average Purchase Value: Total revenue divided by the number of purchases.
Purchase Frequency: Number of purchases divided by the number of unique customers.
Customer Lifespan: Average number of years a customer continues to buy from the business.

Code:

```
import pandas as pd
from datetime import datetime
# Sample customer data
customer_data = {
'customer_id': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
'first_name': ['Alex', 'Sarah', 'Michael', 'Emily', 'David', 'Linda', 'Robert', 'Maria', 'James', 'Anna'],
'last_name': ['Johnson', 'Davis', 'Brown', 'Wilson', 'Lee', 'Martinez', 'White', 'Gonzalez', 'Clark', 'Roberts'],
'first_visit': ['2021-06-15', '2020-03-10', '2019-11-25', '2020-07-22', '2021-02-01', '2020-09-12', '2019-05-18', '2018-08-08', '2021-01-05', '2022-04-15'],
'recent_visit': ['2022-11-30', '2022-12-01', '2022-10-20', '2022-09-15', '2022-08-30', '2022-07-20',
        '2022-06-10', '2022-05-22', '2022-04-30', '2022-03-20']
}
# Sample transaction data
transaction_data = {
'transaction_id': [101, 102, 103, 104, 105, 106, 107, 108, 109, 110],
'customer_id': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
'net_amount': [150, 200, 350, 120, 500, 250, 300, 400, 600, 700],
'total_amount': [135, 190, 315, 108, 450, 225, 270, 380, 540, 630]
}
# Convert dictionaries to DataFrames
customers = pd.DataFrame(customer_data)
transactions = pd.DataFrame(transaction_data)
# Convert date columns to datetime
customers['first_visit'] = pd.to_datetime(customers['first_visit'])
customers['recent_visit'] = pd.to_datetime(customers['recent_visit'])
# Calculate total purchase value, average purchase value, frequency, and recency for each customer
customer_lifetime_value = transactions.groupby('customer_id').agg(
total_purchase_value=('total_amount', 'sum'),
transaction_count=('transaction_id', 'count')
).reset_index()
customer_lifetime_value['average_purchase_value'] = customer_lifetime_value['total_purchase_value'] /
customer_lifetime_value['transaction_count']
customer_lifetime_value['purchase_frequency'] = customer_lifetime_value['transaction_count'] /
customers.shape[0]
customer_lifetime_value['customer_lifespan'] = (customers['recent_visit'] -
        customers['first_visit']).dt.days / 365
# Calculate CLTV
customer_lifetime_value['CLTV'] = customer_lifetime_value['average_purchase_value'] *
customer_lifetime_value['purchase_frequency'] * customer_lifetime_value['customer_lifespan']
# Identify high-value customers (example threshold: CLTV > 500)
high_value_customers = customer_lifetime_value[customer_lifetime_value['CLTV'] > 500]
print(high_value_customers)
```

Modeling: Use machine learning models to predict CLTV more accurately. Common models include Linear Regression, Decision Trees, Random Forests, etc.

Implementation: Implement the chosen model and calculate CLTV for each customer. Identify high-value customers based on predicted CLTV

2. Churn Prediction:

- At-Risk Customers: Michael Brown (ID: 3) and David Lee (ID: 5) have not made recent visits and have lower transaction amounts, indicating a higher risk of churning.

Calculation: First, collect customer and transaction data, ensuring it's clean and organized. Next, prepare the data by creating features like recency (time since last purchase), frequency (number of transactions), and monetary value (total spent). Then, analyze the data to identify patterns. Choose a suitable model (like logistic regression or decision trees) and train it on your data. Split the data into training and testing sets, and evaluate the model using accuracy, precision, and recall. Use the model to predict which customers are likely to churn. Finally, develop strategies to retain at-risk customers, such as personalized offers and targeted marketing campaigns. Regularly monitor and update your model to keep it accurate.

Code:

```
 import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Sample customer data
customer_data = {
    'customer_id': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'first_visit': ['2021-06-15', '2020-03-10', '2019-11-25', '2020-07-22', '2021-02-01', '2020-09-12', '2019-05-18', '2018-08-08', '2021-01-05', '2022-04-15'],
    'recent_visit': ['2022-11-30', '2022-12-01', '2022-10-20', '2022-09-15', '2022-08-30', '2022-07-20', '2022-06-10', '2022-05-22', '2022-04-30', '2022-03-20'],
    'total_amount': [135, 190, 315, 108, 450, 225, 270, 380, 540, 630],
    'status': ['Active', 'Active', 'Inactive', 'Active', 'Inactive', 'Active', 'Active', 'Inactive', 'Active', 'Active']
}

# Convert dictionary to DataFrame
customers = pd.DataFrame(customer_data)

# Convert date columns to datetime
customers['first_visit'] = pd.to_datetime(customers['first_visit'])
customers['recent_visit'] = pd.to_datetime(customers['recent_visit'])

# Calculate features
customers['recency'] = (pd.Timestamp.today() - customers['recent_visit']).dt.days
customers['tenure'] = (customers['recent_visit'] - customers['first_visit']).dt.days
customers['churn'] = customers['status'].apply(lambda x: 1 if x == 'Inactive' else 0)

# Feature selection
features = ['recency', 'tenure', 'total_amount']
```

```
X = customers[features]
y = customers['churn']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict churn
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print(f'Precision: {precision}')
print(f'Recall: {recall}')
print(f'F1 Score: {f1}')

# Predict churn for new customers
new_customers = pd.DataFrame({
    'recency': [30, 180],
    'tenure': [365, 730],
    'total_amount': [500, 200]
})

churn_predictions = model.predict(new_customers)
print(f'Predicted Churn: {churn_predictions}')
```

3. Purchase Pattern Analysis:

- **Preferred Payment Methods**: Cash and credit are the most popular payment methods among the customers, with a slight preference for credit transactions.
- **Effective Discounts**: Percentage discounts are more frequently used and result in higher total amounts, suggesting they are more effective than fixed discounts.

Calculation: To analyze purchase patterns, first collect and clean your transaction data, including payment methods and discount types. Aggregate the data to calculate the total number of transactions and total amount spent for each payment method and discount type. Next, visualize the data to identify which payment methods are most popular and which discount types result in higher spending. This will help you

understand customer preferences and the effectiveness of different discounts, allowing you to tailor your marketing strategies accordingly.

Code:

```
import pandas as pd
import matplotlib.pyplot as plt

# Sample transaction data
transaction_data = {
    'transaction_id': [101, 102, 103, 104, 105, 106, 107, 108, 109, 110],
    'payment_method': ['Credit', 'Cash', 'Debit', 'Credit', 'Cash', 'Debit', 'Credit', 'Cash', 'Debit', 'Credit'],
    'total_amount': [135, 190, 315, 108, 450, 225, 270, 380, 540, 630],
    'discount_type': ['Percentage', 'Fixed', 'Percentage', 'Percentage', 'Percentage', 'Percentage', 'Percentage',
'Fixed', 'Percentage', 'Percentage']
}

# Convert dictionary to DataFrame
transactions = pd.DataFrame(transaction_data)

# Analyze preferred payment methods
payment_method_counts = transactions['payment_method'].value_counts()
payment_method_totals = transactions.groupby('payment_method')['total_amount'].sum()

# Analyze effective discounts
discount_type_counts = transactions['discount_type'].value_counts()
discount_type_totals = transactions.groupby('discount_type')['total_amount'].sum()

# Plot preferred payment methods
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
payment_method_counts.plot(kind='bar', color='skyblue')
plt.title('Preferred Payment Methods')
plt.xlabel('Payment Method')
plt.ylabel('Number of Transactions')

plt.subplot(1, 2, 2)
payment_method_totals.plot(kind='bar', color='orange')
plt.title('Total Amount by Payment Method')
plt.xlabel('Payment Method')
plt.ylabel('Total Amount Spent')

# Plot effective discounts
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
discount_type_counts.plot(kind='bar', color='lightgreen')
plt.title('Effective Discounts')
plt.xlabel('Discount Type')
```

```
plt.ylabel('Number of Transactions')

plt.subplot(1, 2, 2)
discount_type_totals.plot(kind='bar', color='purple')
plt.title('Total Amount by Discount Type')
plt.xlabel('Discount Type')
plt.ylabel('Total Amount Spent')

plt.tight_layout()
plt.show()
```

Challenges and Considerations:

While the potential benefits of AI-based predictive analysis are substantial, there are also several challenges and considerations that organizations must address. Firstly, the dependence on unique data means that there are no shortcuts for firms, and they must continuously work on advancing their digital journeys, including AI. [1] Secondly, the successful implementation of an AI-driven predictive analysis program requires a holistic approach, including identifying the business case, setting up the right data ecosystem, building or acquiring appropriate AI tools, and adapting workflow processes, capabilities, and organizational culture. [1]

Additionally, the paper highlights the fact that few organizations have even entered the big data stage in terms of employee management, where the potential has been articulated often and loudly in adequate decisions. The promise of data analytics is easier to explore in some fields, such as marketing, where the data is more abundant and the use cases are more evident.

Conclusion:

AI and predictive analysis are changing how businesses make decisions. Using the historical data and advanced algorithms, these technologies offer valuable insights about the future. AI helps predictive analysis by automating data processing and finding patterns that humans might miss. This combination allows companies to predict trends, optimize operations, and improve customer experiences. From healthcare to finance, retail to manufacturing, AI-driven predictive analysis has many applications. In summary, AI and predictive analysis drive innovation and efficiency across various industries. As these technologies evolve, their ability to forecast outcomes and support decisions will become even more powerful, making data-driven insights essential for every strategic initiative. By integrating customer and transaction data, the retail was able to gain valuable insights into customer behavior, improve customer segmentation, predict CLTV, and identify potential churn risks. These insights enabled the company to implement targeted marketing strategies, enhance customer retention efforts, and optimize promotional offers.

References:

[1] J. Bughin et al., *"Artificial intelligence: the next digital frontier?,"* Jun. 01, 2017. [Online]. Available: https://apo.org.au/node/210501

[2] J.-P. Cabrera-Sánchez, I. R. de Luna, E. Carvajal-Trujillo, and Á. F. V. Ramos, *"Online Recommendation Systems: Factors Influencing Use in E-Commerce,"* Oct. 26, 2020, Multidisciplinary Digital Publishing Institute. doi: 10.3390/su12218888.

[3] N. Soni, E. K. Sharma, N. Singh, and A. Kapoor, *"Impact of Artificial Intelligence on Businesses: from Research, Innovation, Market Deployment to Future Shifts in Business Models,"* Jan. 01, 2019, Cornell University. doi: 10.48550/arxiv.1905.02092.

[4] A. K. Kordon, *"Applied Artificial Intelligence-Based Systems as Competitive Advantage,"* Aug. 01, 2020. doi: 10.1109/is48319.2020.9200097.

[5] S. Chatterjee, S. K. Ghosh, R. Chaudhuri, and B. Nguyen, *"Are CRM systems ready for AI integration?,"* May 30, 2019, Emerald Publishing Limited. doi: 10.1108/bl-02-2019-0069.

[6] D. G. Harkut and K. Kasat, *"Introductory Chapter: Artificial Intelligence - Challenges and Applications,"* in IntechOpen eBooks, IntechOpen, 2019. doi: 10.5772/intechopen.84624.

[7] T. H. Davenport, *"From analytics to artificial intelligence,"* Jul. 03, 2018, Taylor & Francis. doi: 10.1080/2573234x.2018.1543535.

[8] Edwards, J. and Olavsrud, T., "Predictive analytics: Transforming data into future insights," CIO, 10 February 2021. Available: https://www.cio.com/article/228901/what-is-predictive-analytics-transforming-data-into-future-insights.html

[9] Kaplan, E., "Five Key Trends Shaping The Future Of Predictive Analytics," Forbes, 5 October 2023. Available: https://www.forbes.com/councils/forbestechcouncil/2023/10/05/five-key-trends-shaping-the-future-of-predictive-analytics/

[10] Beasley, K., "Unlocking The Power Of Predictive Analytics With AI," Forbes, 11 August 2021. Available: https://www.forbes.com/councils/forbestechcouncil/2021/08/11/unlocking-the-power-of-predictive-analytics-with-ai/