UNIVERSITY OF LIVERPOOL

EBUS633-Big Data Management

Year: 2024-2025

TOPIC : Big Data Analytics in Retail: Insights and Applications

Prepared by :

Name : Christina Ann Jacob

Student id : 201805307

26th Nov , 2024

TABLE OF CONTENTS

INTRODUCTION

Big data analytics has changed decision-making processes across industries by allowing organizations to use the power of structured and unstructured data.This report focuses on using big data analytics in the retail industry. It shows how structured and unstructured data can solve key business problems, improve performance, and innovation. Section 1 reviews recent studies , discussing methodologies and their applications in predictive analytics. We also use the Random Forest algorithm with real-world data to demonstrate its practical use.Section 2 looks at Sweden's market, focusing on sales, profits and customer trends, especially in the women's clothing sector. It ends with some suggestions and ideas about big data in retail's future. The research combined with real examples links the theory with practical usage.

Section 1

(1a): Critical Review of Contemporary Literature on Predictive Analytics in Retail

Predictive analytics can change retail with better demand forecasts, inventory, and customer insights. This review looks at five important studies that address these issues, highlighting their methods, strengths, and limits.

Accurately predicting customer demand to avoid stockouts or overstocking remains one of the major challenges in retail. (Verma, 2020) addresses this issue and talks about how deep learning models like LSTM and TCN can predict what customers might buy by analyzing structured data from transactions. These models help businesses manage their stock better and create tailored marketing strategies.TCN performed better than RandomForest with an F1 score of 0.85 compared to 0.78. However, since these models only use organized data like transaction records, they miss out on useful insights from unstructured sources like customer feedback.Despite this ,these models make inventory management easier , reduce stock shortages, and make marketing strategies more effective to meet customer needs.

Another major issue is managing sparse sales data, particularly for low-demand products.(Pitkin, Manolopoulou and Ross, 2018) introduced a Bayesian hierarchical model designed to address the difficulty of predicting demand for products with low sales volume. It works well for sparse sales data since it handles zero-inflated data and temporal clustering by leveraging information from different goods. For example, in predicting tablet sales, it was 15% more accurate than traditional methods. Although it only relies on structured data such as sales records and lacks insights from customer feedback or social media, it helps reduce unsold goods and balance inventory, making it useful for managing stock and demand prediction.

To overcome the use of structured data only, (Hossam et al., 2024) came up with a smart retail analytics system that combines sales data and unstructured video surveillance to enhance stock management and customer satisfaction.These are GRU models for demand forecasting and YOLO-V8 for live customer tracking, which gave a 95% accuracy. GRU models have further enhanced the demand forecast with an R^2-score improvement of 2.87%, besides reducing the error by 29.31% compared to traditional models. However, this is an expensive system that is hardly accessible to smaller retailers. Making it cheaper would be the key to making SRAS a staple in retail.

The COVID-19 pandemic showed how weak retail supply chains can be and why better forecasting methods are needed.(Yossiri Adulyasak et al., 2023) talk about the use of predictive analytics in handling the COVID-19 crisis, combining sales data with socio-economic information to identify unusual trends. AI allowed retailers to adapt to sudden changes in customers' behavior: it reduced the forecasting error by 20%, stockouts by 15%, and extra inventory by 10%.However, the study points out that the strategies for managing unstructured data were not explained properly, suggesting a need for better processing methods to maximize the potential of AI in retail.

Retail businesses struggle to understand customer sentiment and behavior due to the huge volume and complexity of unstructured data.(Bilal Abu-Salih et al., 2021) created a model that combines social media data with transaction data to improve predictions. Using sentiment analysis ,the model combines marketing and inventory strategies with customer preferences , getting an accuracy of 97.95% and precision of 98.57% .Because this study used unstructured data to gain customer insights, it was selected. However, the test was conducted on only one case study, hence having limits how widely it can be applied.Such a test would be more valid and of more use if it were tested in more diverse retail environments.

Discussion

These studies show how predictive analytics is changing the retail industry. While structured data is great to predict demand, unstructured data gives deeper insights into customers.While (Verma, 2020) and (Pitkin, Manolopoulou and Ross, 2018) focused only on structured data, (Hossam et al., 2024)and (Bilal Abu-Salih et al., 2021) have highlighted the use of unstructured data also like video footage and social media feedback. However, challenges such as data preprocessing, scalability, and costs remain . Future research should focus on the addition of real-time unstructured data to make better decisions more quickly.

(1b): Methodologies and Method-Problem-Data Relationships in Predictive Analytics

Article 1: "Industry-sensitive Language Modeling for Business"

Journal: European Journal of Operational Research

Methodologies used:

BusinessBERT, introduced by (Borchert et al., 2024), is a deep learning model proposed for NLP applications in the business domain.

The main methods they used:

Deep Learning : Used to develop the BusinessBERT framework, based on the transformer model, to understand complex business terms.

Text Mining: Helps pull out important business insights from a lot of unstructured text.

Transfer Learning : BusinessBERT was trained on over 2 billion business-related words, making it better at understanding business terms.

Sentiment Analysis: This is a way to figure out the feelings expressed in business texts, like whether they are positive or negative.

Problem addressed:

The main problem this study tries to solve is the gap between general language models and what businesses really need. General models often struggle understanding specific business terms, which leads to less accurate findings

Data Type:

The research looks at unstructured data, including text from sources like company websites, management reports, and academic papers about business.

Method-Problem-Data Relationship:

BusinessBERT combines deep learning and NLP to handle unstructured data, which is useful for classification of text, sentiment analysis, and other business-related activities in NLP.Transfer learning helps with decision-making in areas like planning and customer insights. The model's biggest strength is turning messy, unstructured text into clear, useful business information, solving the problem of limited understanding in regular NLP tools.This makes it valuable for businesses looking to extract meaning from messy and diverse text sources.

Article 2: "Deep Reinforcement Learning for Inventory Optimization with Non-Stationary Uncertain Demand"

Journal: European Journal of Operational Research

Methodologies used:

(Dehaybe, Catanzaro and Chevalier, 2024) use Deep Reinforcement Learning (DRL) to create flexible inventory management strategies. Here's how they do it:

Reinforcement Learning (RL): An agent learning the best policies by interacting with an environment.

Deep Learning: Neural networks help the agent understand policies and value functions.

Proximal Policy Optimization (PPO): This method is used to keep the training process stable and effective, especially in complicated and changing situations.

Problems addressed:

The study focuses on the Single-Item Stochastic Lot-Sizing Problem (SISLSP), which deals with unpredictable and changing demand. Traditional models find it difficult to adjust to changes in demand, resulting in poor decision-making.

Data type:

This research is based on structured data, including inventory levels, demand forecasts and simulated orders .

Method-Problem-Data Relationship:

DRL and deep learning combined is great for making quick decisions in inventory management. Using PPO keeps the training process stable , even when the data is very complicated or unpredictable. This method helps the system adjust to changing inventory needs , reducing waste and stockouts and shows how reinforcement learning can make business operations smarter by preparing for future needs.

Comparative Analysis:

Both articles talk about the use of deep learning , though they focus on different forms of data and business problems. BusinessBERT utilizes deep learning in natural language processing to analyze and generate insights from unstructured data, improving the analysis of business texts. However, DRL combines with reinforcement learning to improve decision-making in inventory management by using structured data to address uncertain and changing demand.

Both articles show how deep learning can be used to predict things.BusinessBERT stands out in extracting information from massive amounts of unorganized text, making strategic business choices .DRL uses structured data to solve problems in real time, helping businesses learn and adapt quickly to changes in what people want.

Future works:

Despite their success, both methodologies still have opportunities to overcome certain limitations :

Business BERT: Real-time learning will keep the model stay up to date with the latest business trends and changes in the language.

DRL: This is where using real-world data instead of just simulations may make it even stronger and extend its usability to more industries.

Several combined approaches might use the power of BusinessBERT in analyzing unstructured data and DRL in handling structured data for effective performance. A retail company might use this combination to understand customer feedback while managing its inventory based on demand predictions.

Section 1 (1c): Application of Random Forest Algorithm on a Financial Marketing Dataset

Title: "Bank financial sustainability evaluation: Data envelopment analysis with random forest and Shapley additive explanations"

Journal: European Journal of Operational Research

We used the Bank Marketing Dataset from the UCI Machine Learning Repository to show the Random Forest algorithm widely used to predict customer behavior.The goal is to predict whether a customer will subscribe to a term deposit.

Dataset overview :

Demographics: Age, job, marital status, education.

Financial Details: Account balance, housing loan, personal loan.

Campaign Data: Contact method, number of previous campaigns, days since last contact .

The target variable y is binary, indicating whether the customer subscribed (yes=1) or did not (no=0).

ALGORITHM AND RESULTS :

Using the Random Forest classifier, we trained the model on the preprocessed dataset, splitting the data into 70% for training and 30% for testing.

After training the model and running it on the test set, the classifier achieved an accuracy of 89.98%, as shown in the confusion matrix and classification report:

```
Accuracy of the Random Forest classifier: 0.8998083161309348

Confusion Matrix:
 [[11577   389]
 [  970   628]]

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.97      0.94     11966
           1       0.62      0.39      0.48      1598

    accuracy                           0.90     13564
   macro avg       0.77      0.68      0.71     13564
weighted avg       0.89      0.90      0.89     13564
```

Fig 1: Confusion Matrix of the Random Forest Classifier

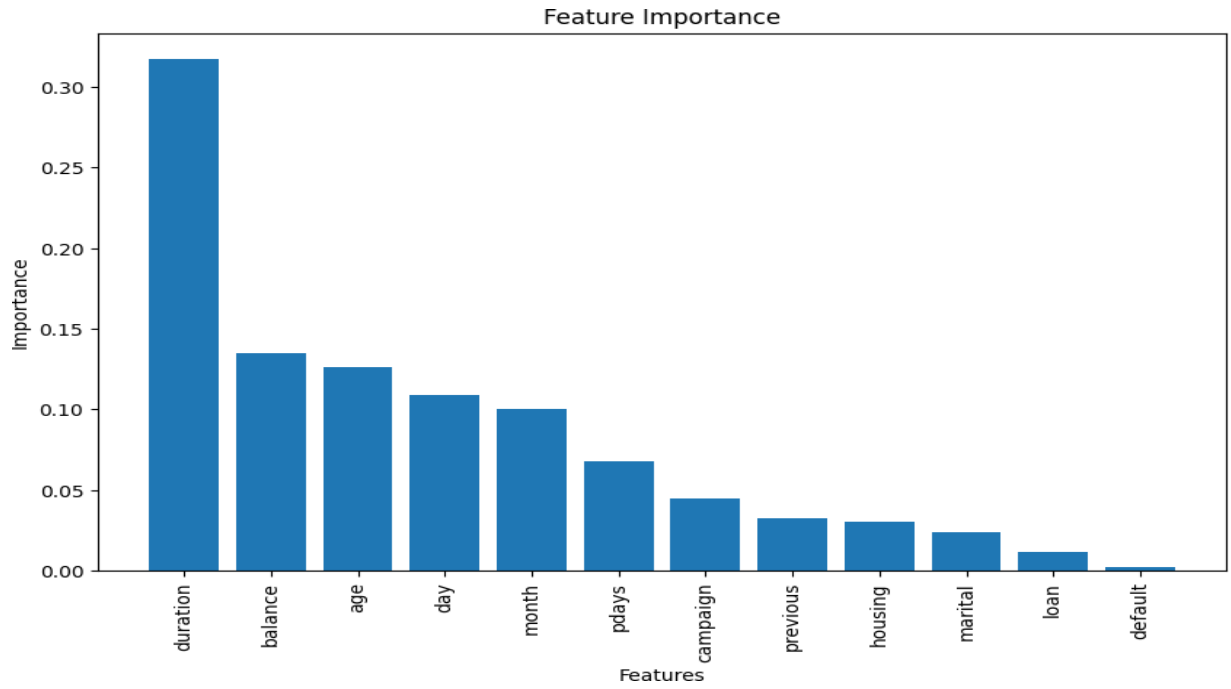

Fig 2: Feature Importance of the Random Forest Classifier

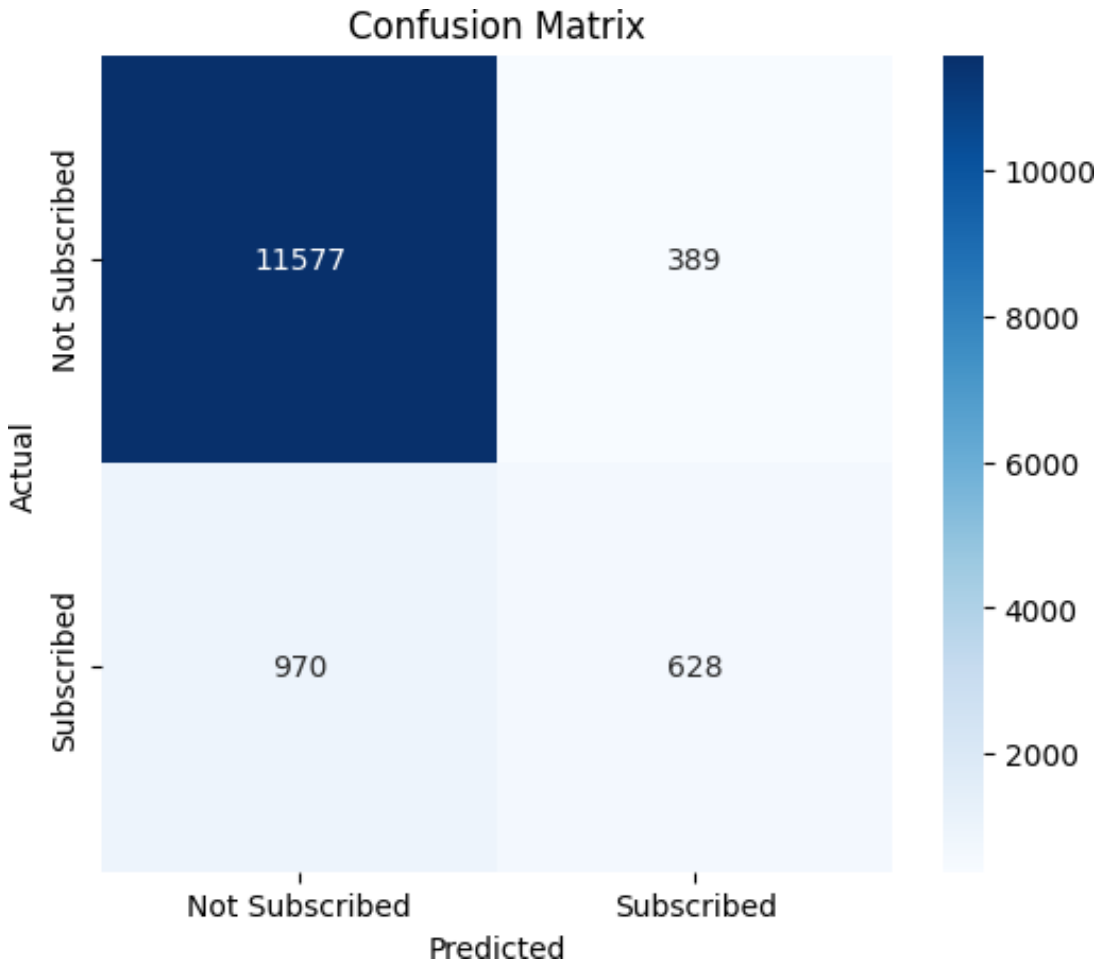

Fig 3: Heatmap Visualization of the Confusion Matrix

RESULTS AND DISCUSSION

The model achieved an overall accuracy of 89.98%, showing strong predictive performance.

The confusion matrix (Fig 1) tells us:

Class 0 (Non-subscribers):The model performs really well , with high precision (92%), recall (97%), and F1-score (94%). This means it's great at identifying non-subscribers and makes very few mistakes.

Class 1 (Subscribers): The model isn't as good at detecting subscribers, which is the minority . Precision is 62%, recall is 39%, and the F1-score is 48%, showing that it often misclassified subscribers due to imbalance in class representation.

Feature importance analysis (Fig 2) shows duration (length of the last contact) is the most important factor, followed by balance, age, and pdays.Duration and balance are the main factors, meaning longer, meaningful customer interactions and financial stability increases subscription rates.

The heatmap (Fig 3) shows the model's difficulty in identifying subscribers, with a high number of false negatives. This means it's missing potential customers who could be targeted better.

Business Implications:

- Focus on customers with higher balance and longer durations for follow-up campaigns.

-Design campaigns addressed to various categories of customers, to increase

conversion rates.

- Removing false negatives can increase resource utilization and an increase in term deposit subscriptions.

Limitation and Future Work

-The dataset is uneven, making it harder for the model to identify the smaller group. Balancing the data can help improve its performance.

-The application in real-time may test the adaptability of the model in changing situations.

-Incorporating unstructured data will add more insights into customer behavior, such as sentiment analysis or text data.

Appendix: The code for data preprocessing, model training, and evaluation is provided in the appendix for reference.

Section 2

(2a): Revenue, Margin, and Number of Customers in Sweden

Last digit of student id : 7

The analysis for Sweden shows that 3 distinct customers can each bring in a total revenue of 57,162.01 units and a margin of 22.24%. This means the business concentrates its sales on a specific group of customers who each bring in a lot of revenue. The strong profit margin shows effective cost management and efficient operations.

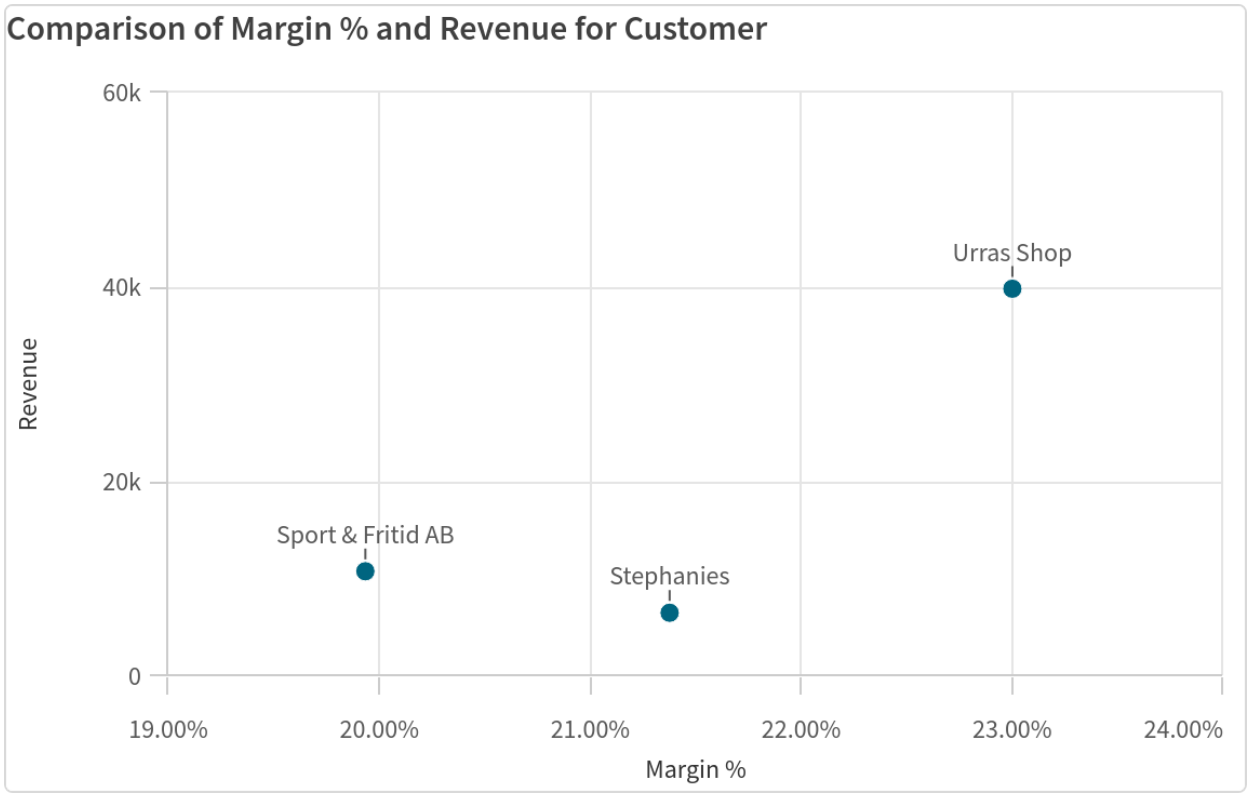

Fig 4: Comparison of Margin% and Revenue for Customer for Sweden

Section 2 (2b): Women’s Clothing Market in Sweden

The women's clothing market in Sweden heavily depends on high-end products, such as the "Really Expensive Coat," which brings in over 10K in revenue and a profit margin of 30%. Mid-segment products, such as the "Party Dress," have potential but need an improved marketing and pricing strategy. This product range can be expanded to achieve better growth.

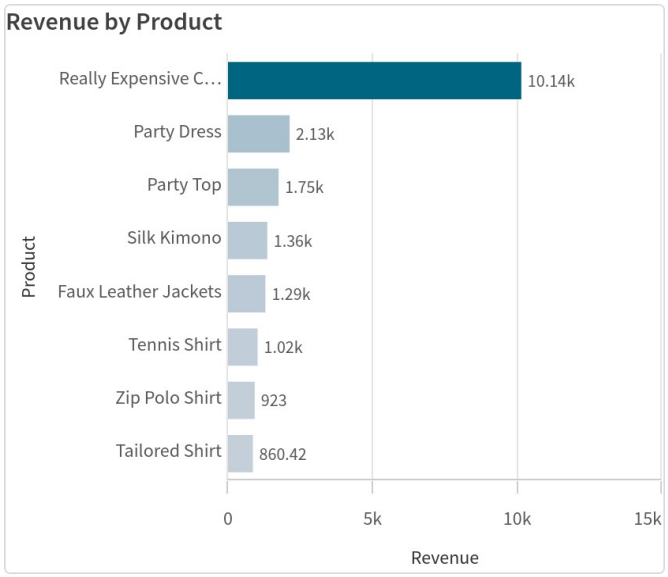

Fig 5: Revenue by product for Sweden

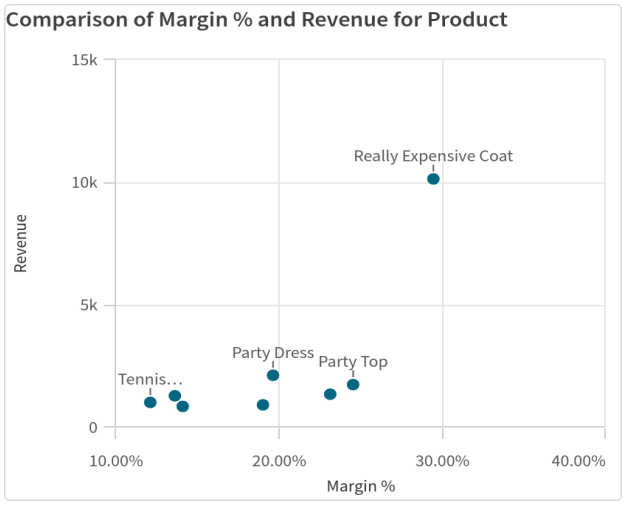

Fig 6: Comparison of Margin % and revenue for product for Sweden

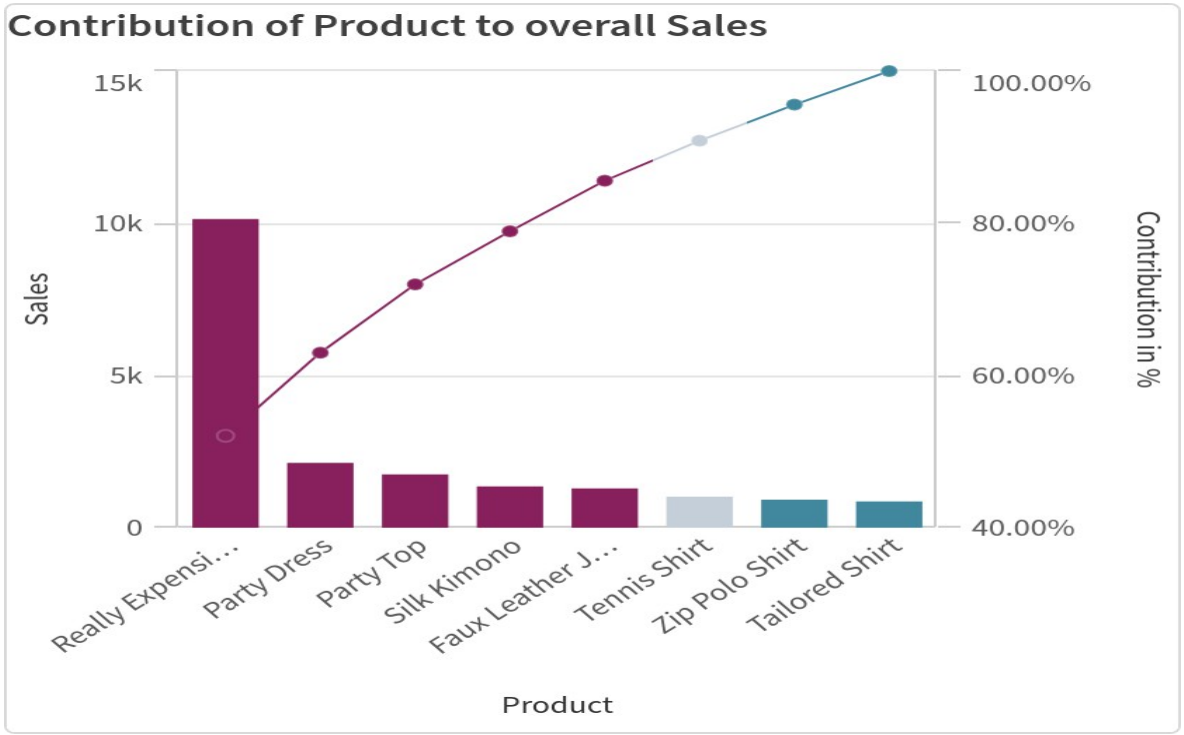

Fig 7 :Contribution of product to overall sales

Executive summary

This report discusses how big data analytics is changing the retail industry thereby making it easier to predict demand, manage inventory, and understand customers. It reviews some important studies and shows how tools like Random Forest, GRU models, and BusinessBERT can analyze both structured data and unstructured data . Even though some are expensive for smaller retailers to use , they help businesses improve decisions . The Random Forest algorithm gave 90% accuracy in predicting customer behavior, showing that time spent with customers and financial stability were the most important features . It also highlights challenges like unbalanced data and why real time adjustments are needed .In conclusion, big data analytics has potential for improvement in retail . But companies need to focus on using real-time data and making these tools affordable for everyone to actually have a big impact .

References

- Bilal Abu-Salih, Pornpit Wongthongtham, Zhu, D., Kit Yan Chan and Amit Rudra (2021). Predictive Analytics Using Social Big Data and Machine Learning. *Springer eBooks*, [online] pp.113–143. doi:https://doi.org/10.1007/978-981-33-6652-7_5.
- Borchert, P., Kristof Coussement, Jochen De Weerdt and Arno De Caigny (2024). Industry-sensitive language modeling for business. *European journal of operational research*, 315(2). doi:https://doi.org/10.1016/j.ejor.2024.01.023.
- Dehaybe, H., Catanzaro, D. and Chevalier, P. (2024). Deep Reinforcement Learning for inventory optimization with non-stationary uncertain demand. *European Journal of Operational Research*, 314(2), pp.433–445. doi:https://doi.org/10.1016/j.ejor.2023.10.007.
- Hossam, A., Ramadan, A., Magdy, M., Abdelwahab, R., Ashraf, S. and Mohamed, Z. (2024). Revolutionizing Retail Analytics: Advancing Inventory and Customer Insight with AI. *arXiv (Cornell University)*. [online] doi:https://doi.org/10.48550/arxiv.2405.00023.
- Moro, S., Rita, P. and Cortez, P. (2012). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/222/bank+marketing [Accessed 26 Nov. 2024].
- Pitkin, J., Manolopoulou, I. and Ross, G. (2018). *Bayesian hierarchical modelling of sparse count processes in retail analytics*. [online] arXiv.org. Available at: https://arxiv.org/abs/1805.05657 [Accessed 20 Nov. 2024].
- Shi, Y., Charles, V. and Zhu, J. (2024). Bank financial sustainability evaluation: Data envelopment analysis with random forest and Shapley additive explanations. *European Journal of Operational Research*, [online] 321(2), pp.614–630. doi:https://doi.org/10.1016/j.ejor.2024.09.030.

- Verma, A. (2020). *Consumer Behaviour in Retail: Next Logical Purchase using Deep Neural Network*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2010.06952.
- Yossiri Adulyasak, Cohen, M.C., Warut Khern-am-nuai and Krause, M. (2023). Retail Analytics in the New Normal: The Influence of Artificial Intelligence and the Covid-19 Pandemic. *IEEE Engineering Management Review*, pp.1–26. doi:https://doi.org/10.1109/emr.2023.3337415.

Appendix: Code for Data Preprocessing, Model Training, and Evaluation

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, classification_report,
confusion_matrix

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns



df = pd.read_csv('bank-full.csv', sep=';')


# Display the first few rows and data types


print(df.head())

print(df.dtypes)
```

```
# Handle missing or 'unknown' values
df = df.replace('unknown', pd.NA)
df = df.dropna(axis=1, how='any')  # Drop columns with unknown values
df = df.dropna(axis=0, how='any')  # Drop rows with unknown values



print(df.head())


# Encode categorical variables using LabelEncoder
label_encoders = {}
categorical_columns = df.select_dtypes(include=['object']).columns


for col in categorical_columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
print(df.head())


# Separate features (X) and target variable (y)
```

```
X = df.drop('y', axis=1)  # 'y' is the target variable
y = df['y']

# Split the dataset into training and testing sets (70% training, 30% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Scale the features using StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

print("Data preprocessing complete.\n")

# Initialize and train the Random Forest classifier
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
rf_classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = rf_classifier.predict(X_test)
```

```
# Evaluate the model's performance
acc = accuracy_score(y_test, y_pred)
conf_mat = confusion_matrix(y_test, y_pred)
class_rep = classification_report(y_test, y_pred)

print("Accuracy of the Random Forest classifier:", acc)
print("\nConfusion Matrix:\n", conf_mat)
print("\nClassification Report:\n", class_rep)

# Plot feature importance
feature_importances = rf_classifier.feature_importances_
features = X.columns

plt.figure(figsize=(10, 6))
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), feature_importances[indices], align="center")
plt.xticks(range(X.shape[1]), [features[i] for i in indices], rotation=90)
```

```
plt.xlabel("Features")
plt.ylabel("Importance")
plt.tight_layout()
plt.show()

# Plot confusion matrix as a heatmap
plt.figure(figsize=(6, 5))
sns.heatmap(confusion_mat, annot=True, fmt="d", cmap="Blues",
            xticklabels=["Not Subscribed", "Subscribed"],
            yticklabels=["Not Subscribed", "Subscribed"])
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.tight_layout()
plt.show()
```