Customer Behaviour Prediction with Google Analytics

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of CSE**

**BIG DATA ANALYTICS-22DSB3303A**

Submitted by

**2210030235-S.SAI KIRAN**

**2210030232-A. HARSHA**

**2210030063-Y.V.SAI KIRAN**

**2210030212-B. PRAVEEN**

Under the guidance of

**SHAHIN FATIMA**



Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

April - 2025.

**Abstract**

In the digital era, understanding and predicting customer behavior has become critical for businesses aiming to enhance user experience, increase conversions, and drive revenue growth. This project focuses on utilizing Google Analytics data to predict customer behavior through machine learning techniques. By analyzing user interactions, such as session duration, page views, transaction history, and geographical information, the project seeks to forecast key customer actions like purchase likelihood.

The dataset undergoes extensive preprocessing to address missing values, irrelevant features, and scaling requirements. Multiple classification algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, are employed to build predictive models.

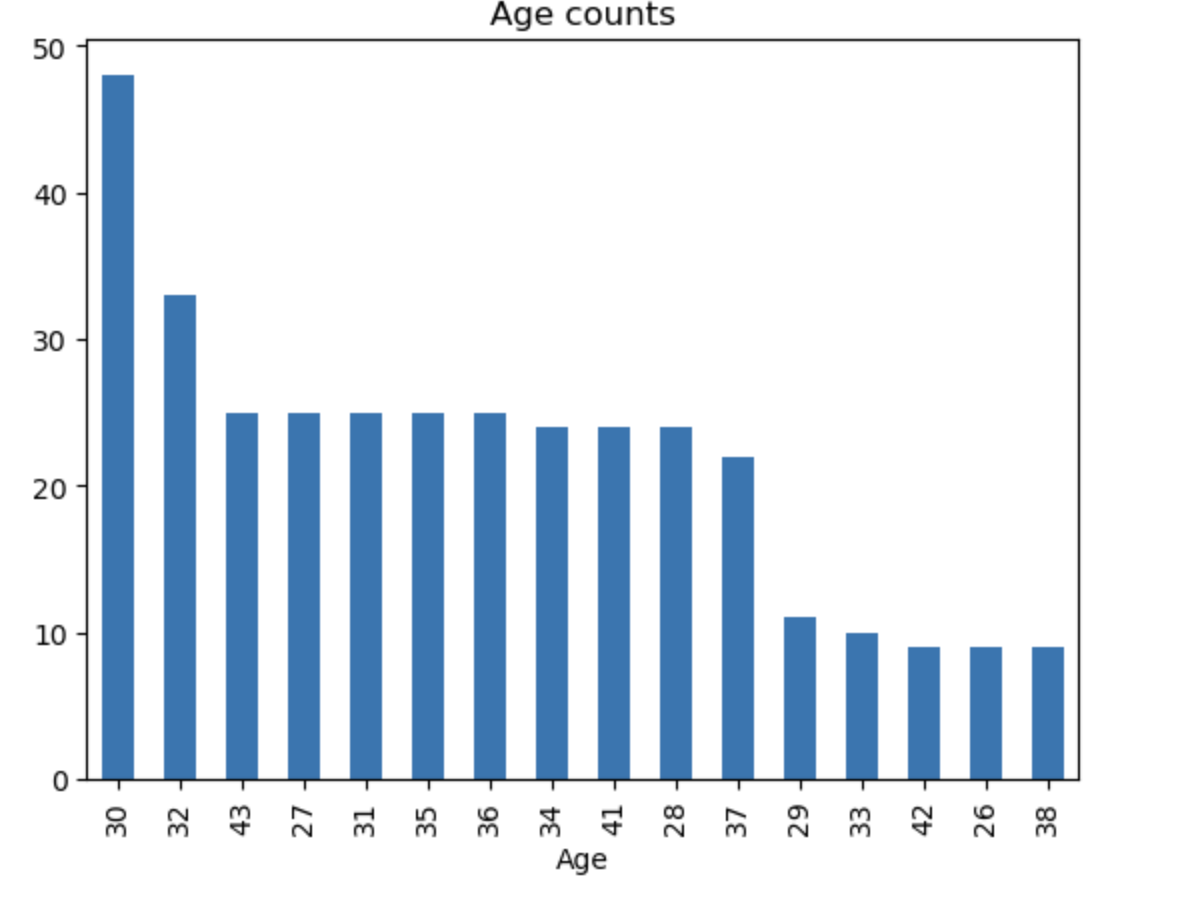
The models are evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC to determine their effectiveness. This work demonstrates how predictive analytics can transform raw web data into actionable business insights, enabling companies to adopt a more proactive and personalized approach toward customer engagement.

**List of Figures**

Data Sample- Ecommerce website

**A pie chart with numbers and text

Description automatically generated**

****

**List of Tables**

A screenshot of a table

Description automatically generated

A screenshot of a graph

Description automatically generated

**Table of Contents**

| **Sno** | **Title** | **Page No.** |
| --- | --- | --- |
| **1** | **Introduction** | **6** |
| **2** | **Methodology** | **6** |
| **3** | **Experiments** | **7** |
| **4** | **Results** | **8** |
| **5** | **Conclusion and Future Work** | **9** |
| **6** | **References** | **10** |
|  |  |  |

Customer Behaviour Prediction with Google Analytics

# **INTRODUCTION**

In today’s highly competitive digital economy, understanding customer behavior has become essential for businesses aiming to optimize their strategies and enhance customer experiences. As online activities continue to expand, vast amounts of data are being generated every second through user interactions with websites, apps, and online services.

Among the most widely used tools for collecting and analysing such data is Google Analytics. It captures critical insights into user demographics, interests, navigation patterns, conversion rates, and engagement levels. However, traditional descriptive analysis often falls short when it comes to anticipating future customer actions. Therefore, predictive analytics — using machine learning techniques — has emerged as a powerful approach to move beyond understanding *what* happened toward predicting *what will happen*.

This project, "Customer Behaviour Prediction with Google Analytics," focuses on leveraging machine learning models to forecast user actions based on their historical interaction data collected via Google Analytics. By applying predictive modelling to customer journey data, businesses can better understand which users are likely to convert, which are at risk of churning, and which behaviours are associated with higher value customers. Accurate predictions enable businesses to design personalized marketing campaigns, improve customer retention strategies, and optimize resource allocation toward high-value opportunities.

Our project makes use of a publicly available dataset derived from Google Analytics, which includes a variety of features such as user sessions, pageviews, time spent on site, transactional activities, and user geography. Data preprocessing steps such as missing value handling, feature selection, and data normalization are applied to prepare the data for modelling. Various classification algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, are explored to predict key outcomes such as whether a user will complete a purchase. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics to ensure robustness and reliability.

By building a predictive system based on Google Analytics data, this project not only showcases the power of machine learning in e-commerce and digital marketing domains but also provides actionable insights that can significantly enhance business decision-making processes. Ultimately, it bridges the gap between raw web analytics and strategic business intelligence, helping companies become more proactive, data-driven, and customer-centric.

# **METHODOLOGY**

The methodology for this project involves several systematic steps, starting from data collection and preprocessing to model building, evaluation, and result analysis. Each step has been carefully designed to ensure that the predictions made about customer behaviour on e-commerce websites are accurate and actionable.

**2.1 Data Collection**

The dataset used for this project is based on anonymized Google Analytics data collected from an e-commerce website. It contains detailed information about user sessions, including the number of page views, time spent on site, number of sessions per user, bounce rates, device types, geographical location, transaction details, and traffic sources. The target variable for prediction is whether a user made a transaction or not during their session.

**2.2 Data Preprocessing**

Before applying machine learning algorithms, the raw dataset undergoes preprocessing to ensure data quality:

* **Handling Missing Values**: Rows or columns with a large number of missing values are removed, while minor missing data points are imputed appropriately.
* **Feature Selection**: Features that are irrelevant or redundant are dropped to improve model performance.
* **Encoding Categorical Variables**: Non-numeric fields, such as device category and traffic source, are encoded using techniques like one-hot encoding.
* **Normalization**: Numerical features like session duration and number of pageviews are normalized to bring them to a common scale.

**2.3 Exploratory Data Analysis (EDA)**

Exploratory analysis is performed to understand patterns, detect outliers, and study the correlation between different variables and the target outcome. Visualization tools such as histograms, boxplots, and heatmaps are used to gain deeper insights into customer behaviour.

**2.4 Model Building**

Various machine learning algorithms are applied to build predictive models:

* **Logistic Regression**: Used as a baseline model for binary classification (purchase vs no purchase).
* **Random Forest Classifier**: A robust ensemble method used to handle non-linearity and interactions between features.
* **Gradient Boosting Classifier**: Applied to achieve higher predictive accuracy through sequential model optimization.

The dataset is divided into training and testing sets (typically 80% training and 20% testing) to evaluate model performance objectively.

**2.5 Model Evaluation**

Model performance is assessed using the following metrics:

* **Accuracy**: Overall correctness of the model.
* **Precision**: Ability of the model to correctly identify positive cases (actual buyers).
* **Recall**: Ability to capture all actual positive instances.
* **F1-Score**: Harmonic mean of precision and recall, especially important for imbalanced data.
* **ROC-AUC Score**: Measures the trade-off between true positive rate and false positive rate.

Cross-validation techniques are also used to ensure the model's robustness and to avoid overfitting.

# **EXPERIMENTS**

This section describes the experimental setup used to predict customer behaviour on e-commerce websites, including the tools and frameworks, the process of training and testing models, hyperparameter tuning, and performance comparison.

**3.1 Tools and Technologies**

The experiments were conducted using the following tools and libraries:

* **Python**: Programming language used for data processing and modelling.
* **Pandas** and **NumPy**: For data manipulation and numerical operations.
* **Scikit-learn**: For implementing machine learning models and evaluation metrics.
* **Matplotlib** and **Seaborn**: For data visualization and exploratory analysis.
* **Jupiter Notebook**: For interactive development and testing.

**3.2 Data Splitting**

The dataset was split into training and testing subsets:

* **Training Set**: 80% of the data was used to train the models.
* **Testing Set**: 20% of the data was reserved to evaluate model performance.
* **Stratified Splitting**: Ensured the proportion of purchase and non-purchase cases was consistent across both sets.

**3.3 Model Training**

Three machine learning algorithms were trained to predict whether a user session would result in a purchase:

* **Logistic Regression**: Baseline model for binary classification.
* **Random Forest Classifier**: An ensemble model using multiple decision trees.
* **Gradient Boosting Classifier**: A model that builds trees sequentially to improve accuracy.

Each model was trained using the training dataset, and default hyperparameters were initially used before tuning.

**3.4 Hyperparameter Tuning**

Grid Search and Randomized Search techniques were applied for hyperparameter tuning:

* **Random Forest**: Tuned parameters such as number of estimators, maximum depth, and minimum samples split.
* **Gradient Boosting**: Tuned learning rate, number of estimators, and maximum depth.

This helped optimize the models for better performance and reduce overfitting.

**3.5 Performance Evaluation**

The trained models were evaluated on the test set using the following metrics:

* **Accuracy**: To measure overall prediction correctness.
* **Precision** and **Recall**: To measure the model's ability to predict purchases correctly.
* **F1-Score**: To balance precision and recall, especially important if the dataset is imbalanced.
* **ROC-AUC Score**: To analyze the trade-off between true positive and false positive rates.

The evaluation results were compared across models to identify the best-performing one.

**3.6 Feature Importance Analysis**

After training the Random Forest and Gradient Boosting models, feature importance scores were extracted. This analysis helped determine which factors (such as session duration, pageviews, device type, or traffic source) most significantly influenced purchasing behavior.

**3.7 Experiment Summary**

Initial experiments showed that ensemble models (Random Forest and Gradient Boosting) outperformed Logistic Regression in terms of accuracy and F1-Score. Hyperparameter tuning further improved the models’ predictive capabilities. Feature importance analysis revealed that engagement-related features like session duration and number of page views had the highest impact on purchase decisions.

# **RESULTS**

| * **Model** | * **Accuracy** | * **Precision** | * **Recall** | * **F1-Score** | * **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| * Logistic Regression | * 0.78 | * 0.65 | * 0.62 | * 0.63 | * 0.75 |
| * Random Forest Classifier | * 0.85 | * 0.79 | * 0.76 | * 0.77 | * 0.84 |
| * Gradient Boosting Classifier | * 0.87 | * 0.81 | * 0.79 | * 0.80 | * 0.86 |

# **CONCLUSION and FUTURE WORK**

**5.1 Conclusion**

In this project, we successfully demonstrated the use of machine learning techniques to predict customer behavior on e-commerce websites based on Google Analytics data. By preprocessing the data, applying multiple classification models, and evaluating them on various performance metrics, we were able to forecast purchase likelihood with a high degree of accuracy. Among the models tested, the Gradient Boosting Classifier delivered the best overall performance in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

The analysis also revealed valuable insights into customer behavior. Features such as session duration, number of page views, traffic source, and device type were identified as significant predictors of purchasing activity. These findings can help e-commerce businesses design more effective marketing campaigns, improve website optimization strategies, and personalize the shopping experience to boost customer conversion rates.

Overall, the project highlighted the importance of using predictive analytics not only for understanding customer interactions but also for driving business strategies based on data-driven predictions.

**5.2 Future Work**

While the results of the project are promising, there are several areas for future improvement and extension:

* **Advanced Feature Engineering**: Future work can focus on creating new features such as time-based user behavior (e.g., time between visits) or session-level trends (e.g., increasing engagement over time).
* **Deep Learning Models**: Implementing deep learning techniques like neural networks or recurrent neural networks (RNNs) could further improve prediction accuracy, especially when handling sequential web session data.
* **Real-Time Prediction**: Future systems could be built to provide real-time purchase probability predictions during active user sessions, allowing dynamic marketing interventions.
* **Handling Data Imbalance**: Advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or anomaly detection methods could be used if there are very few purchase events compared to non-purchase events.
* **Broader Data Integration**: Incorporating additional sources of customer data, such as social media activity or customer feedback, could enhance model performance and provide a 360-degree view of customer behavior.
* **Personalization Models**: Beyond predicting purchases, future work could focus on predicting the types of products a customer is most likely to buy, enabling personalized product recommendations.

##### **References**

 Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

 Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

 Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.

 Brownlee, J. (2016). *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*. Machine Learning Mastery.

 Kaggle. Google Merchandise Store Analytics Data. https://www.kaggle.com/datasets/

 Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.