Team Members

Vaneesha S Kumar (221166)
Nikhil Gupta (220708)
Nischay Patel (220721)

CGS616 - Assignment 1B

OVERVIEW

To simulate the behavioral analysis of the Online shoppers purchasing intention dataset, we developed the Drift diffusion model from scratch due to the constraints in implementing existing libraries. We modelled the population intention for the people who purchased the product and identified the mean and standard deviation for their Response times. To enhance the fitting of DDM on the data we estimated the parameters (Drift rate, Decision boundary and Gaussian noise) using linear estimation and trial calling. Additionally, we analysed the variation of drift rates for a few specific "traffic types" and identified the traffic medium which ended with the quickest purchase.

METHODOLOGY

Exploratory Data Analysis:

The Online shoppers purchasing intention dataset has very few missing values and all features of the dataset are relevant to the purchasing intention based on inference. We visualised the behaviour of features with respect to the "Revenue" using various plots.

Feature Selection:

Based on the visualisations, we selected the following feature which favoured [Revenue=True] [exit_rates, Page_value, Bounce_rate, traffic_type , visitor_type]

Feature scaling and Drift rates estimation:

We used a linear regression to predict the drift rate (A_t) based on the selected features.
Eqn: (A_t) = A1*exit_rates + A2*page_value + A3*bounce_rate + A4*traffic_type + A5*visitor_type + intercept

The model was trained on the dataset we used, and the learned weights *(A1, A2, A3, A4 and A5)* and intercept were extracted. The predicted drift rates for each time step were computed and displayed for inspection, and optionally, added to the original data frame for further analysis.

Drift Diffusion Model Simulation:

We wrote the code using the fundamental equation *dx= Adt+ cdW* to simulate decision-making processes, where the drift is dynamically estimated based on factors like *exit_rates, Page_value, Bounce_rate, traffic_type , visitor_type*. The model simulates multiple trials, updating evidence over time with noise, and tracks whether the decision reaches a threshold for purchase or not. This generates the simulated graph for all the rows by changing the num_trials parameter.

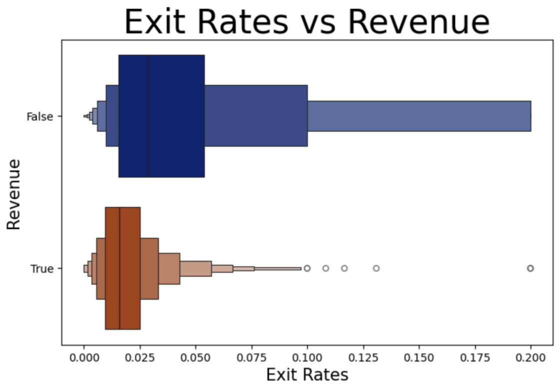
We computed mean and standard deviation of the reaction time and plotted Frequency vs Reaction Time Distribution.

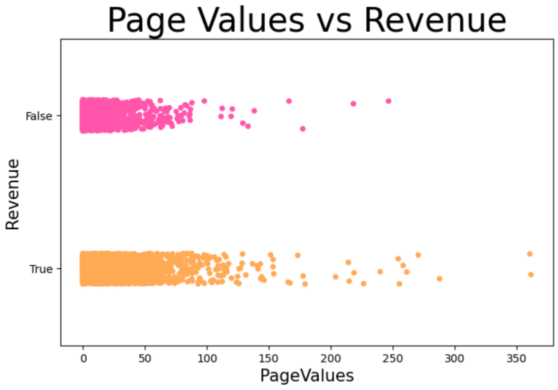
Traffic Type analysis:

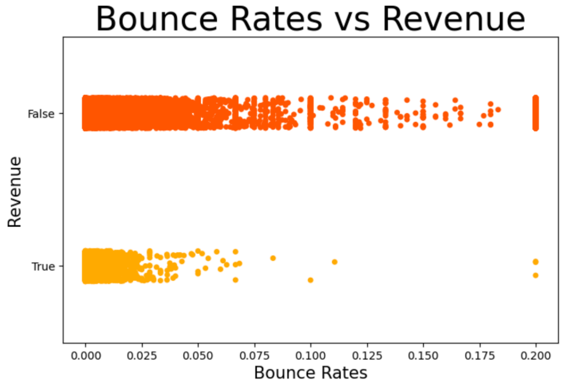
Using the simulation and linear regression, we analysed drift rates for all 20 traffic types and observed the traffic type with the quickest purchase.

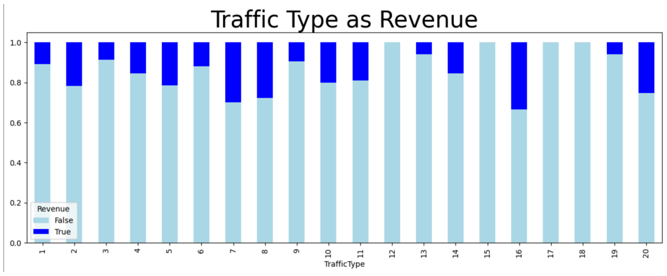
RESULTS AND DISCUSSION

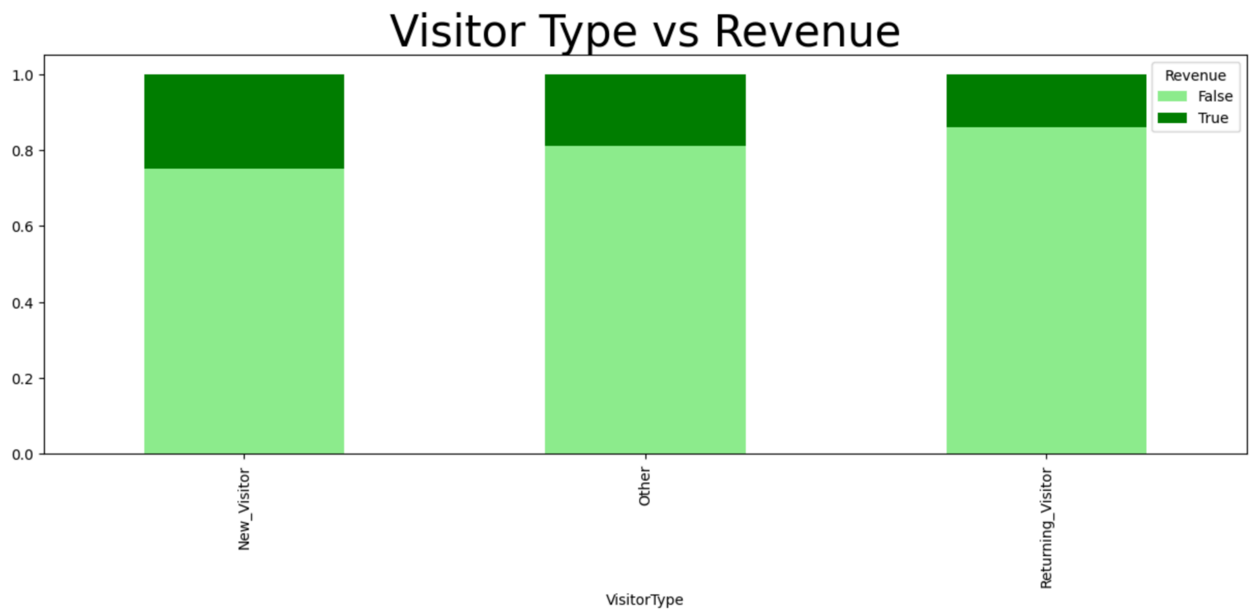
Feature Selection:











All features of the dataset were plotted wrt revenue to observe which features the target variable (Revenue) depended on. After a thorough analysis, it was observed that ExitRates, BounceRates, PageValues, TrafficType and VisitorType were the required features.

RATIONALE FOR PARAMETER FINALIZATION

Decision Boundary:

The Decision Boundary was set to 1 for positive boundary and -0.8 for negative boundary.

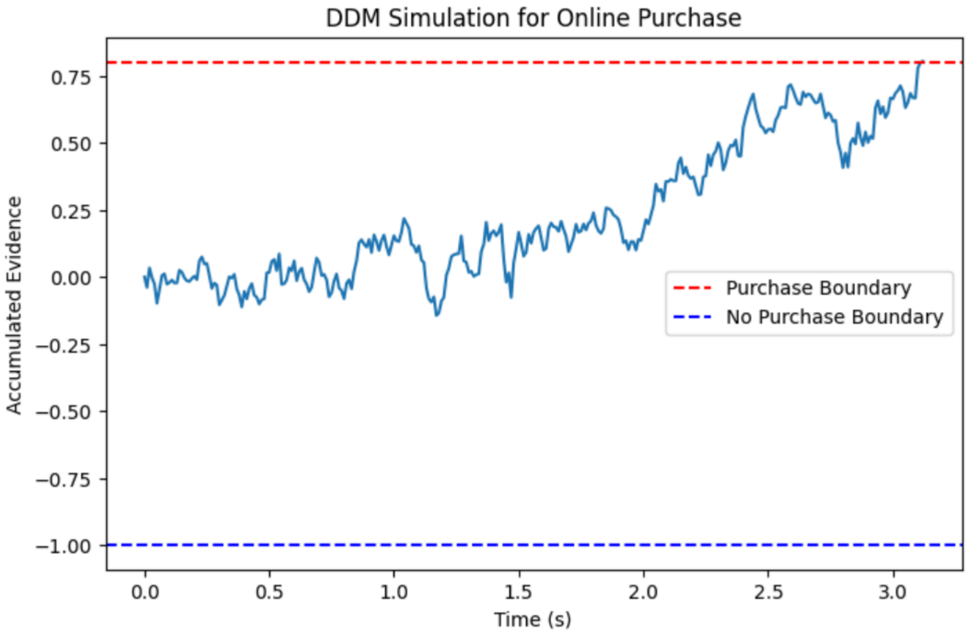
- The linear regression model learns a function for A_t (drift rate) based on historical data.
- Since only 6.6% of users purchased, the model is mostly trained on non-purchasers (93.4%).
- As a result, the learned drift rates A_t will be biased towards predicting no-purchase.
- Most samples will have a negative or small A_t, meaning the accumulated evidence will more often go toward −B (no purchase).
- Fewer users will have high A_t, making it harder to reach +B (purchase).
- This means purchases will be predicted less frequently, even if a user has high "purchase-intent" features.
- For most users, evidence accumulation will move quickly to −B (no purchase), meaning short reaction times.
- For the small number of purchasing users, evidence accumulation will take longer to reach +B, leading to longer reaction times for purchases.

To ensure that the drift diffusion model models this trend, the decision boundary was adjusted to the above mentioned values.

Starting Point:
Given that only 6.6% of users purchased the product, an asymmetric starting point (e.g., closer to the no-purchase boundary) might lead to premature terminations.
By setting it at 0, users have a fair chance to accumulate evidence in either direction without an artificial bias toward no-purchase.

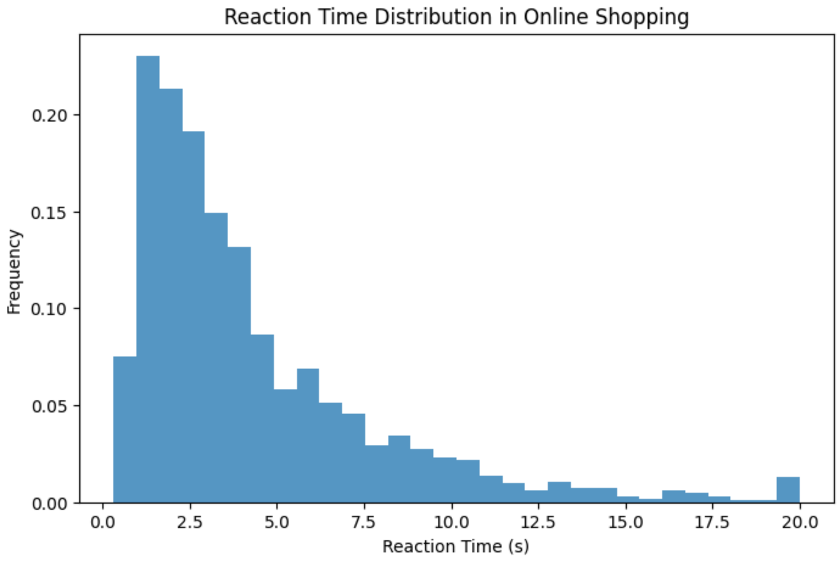
Drift Rate:
A linear regression model was trained on features like *exit_rates, Page_value, Bounce_rate, traffic_type* , *visitor_type* to predict the drift rate for each user.
This means A_t is a function of real-world behavioral factors, rather than being arbitrarily chosen.

MEAN AND SD OF REACTION TIME FOR THE PURCHASE (REVENUE = TRUE)



This is an example graph for evidence accumulation for one user. The reaction time is around 3s.
The mean of reaction times was found to be **4.292699161425577s.**
The standard deviation of reaction times was found to be **3.539405856235836s.**

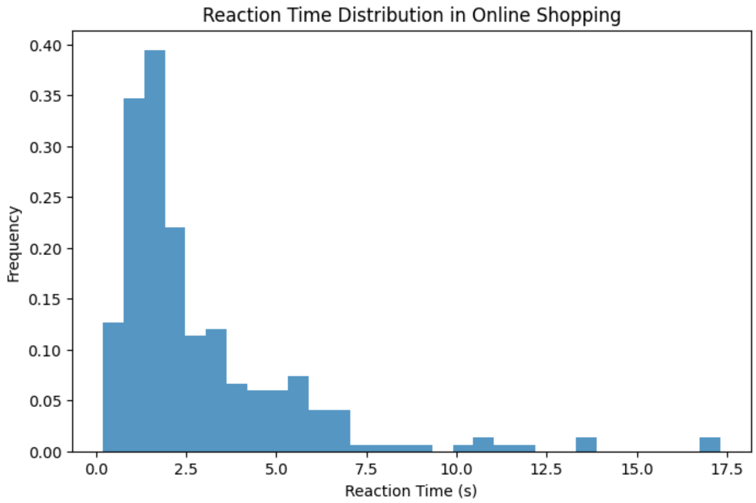
The figure below shows the distribution of reaction times of all users. The distribution follows the expected trends as it is initially **right skewed** with a long tail at the end.



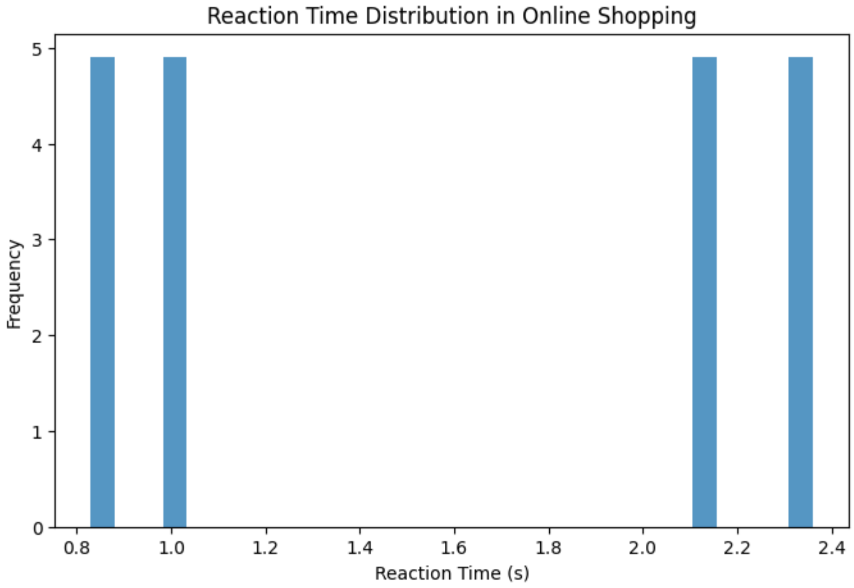
TRAFFIC TYPE ANALYSIS

Using the simulation and linear estimation, we analysed drift rates for all 20 traffic types and found out the traffic type with the quickest purchase (reaction time = 0.19).

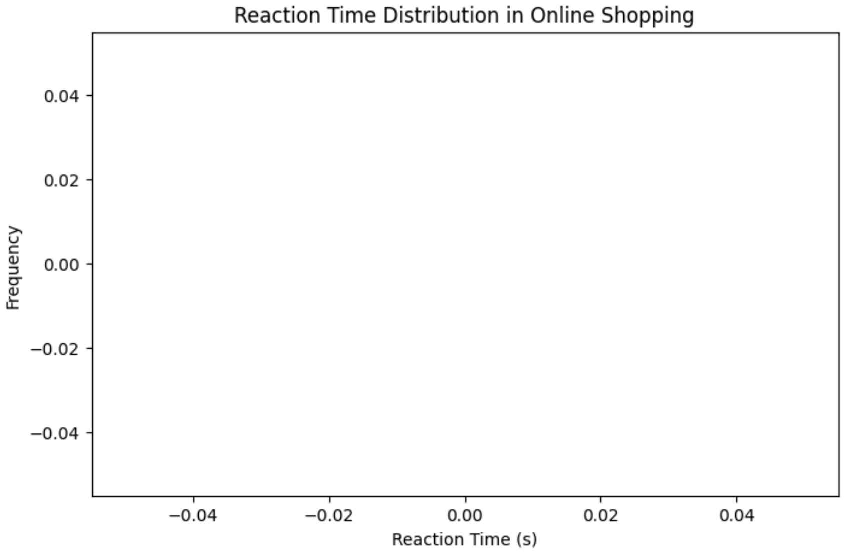
- We obtain some roughly right skewed graphs for reaction time where there is enough data for plotting the graph.
- Some traffic types have very less or no data for purchase. For traffic types with less data points, we obtain a comparatively very high or very low value of reaction time due to fewer and extreme values.



Minimum reaction time for traffic type 1: 0.19s (lowest of all)
Mean drift rate for traffic type 1: 3.580963740458015



Minimum reaction time for traffic type 9: 0.8300000000000001 (very high value due to less data)
Mean drift rate for traffic type 9: 1.9874999999999998



No Purchase in traffic type 12 (no data).

LIMITATIONS

Limited Data for Each Traffic Type:

Due to the limited data available, the response time graph lacks smoothness, resulting in a rough right-skewed plot.

Implementing the DDM Model from Scratch:

As the Hddm library is deprecated and the PyDDM library did not offer a model compatible with our dataset, we had to define and implement the model ourselves.

REFERENCES

[1]“Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks | Semantic Scholar.” Accessed: Jan. 26, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Real-time-prediction-of-online-shoppers%E2%80%99-purchasing-Sakar-Polat/747e098f85ca2d20afd6313b11242c0c427e6fb3

[2]C. E. Myers, A. Interian, and A. A. Moustafa, “A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences,” Front. Psychol., vol. 13, Dec. 2022, doi: 10.3389/fpsyg.2022.1039172.

[3]S. Lalvani and A. Katsaggelos, “Crowdsourcing with the drift diffusion model of decision making,” Sci. Rep., vol. 14, no. 1, p. 11311, May 2024, doi: 10.1038/s41598-024-61687-y.

[4]“PyDDM Cookbook — PyDDM 0.8.1 documentation.” Accessed: Jan. 26, 2025. [Online]. Available: https://pyddm.readthedocs.io/en/latest/cookbook/index.html#model-components

CODE FILES:

CGS616_1B.ipynb

CONTRIBUTIONS:

We all worked equally on all aspects of the Assignment 1A and 1B, including coding, report writing and analysis of the results
Particularly for Assignment 1A, we analyzed and interpreted data individually on the following topics:
Vaneesha S. Kumar: “Abortion Rights”
Nischay Patel: “LGBTQ+ Rights”
Nikhil Gupta: “Gun Ownership”