

LOGISTIC REGRESSION MODEL



Overview: Logistic regression is a classic and interpretable statistical method commonly used for binary classification. It was selected for its simplicity, transparency and ability to provide clear insights regarding the influence of individual features on the target outcome.

Training Method: The model was trained using 10-fold cross-validation. In each iteration, the model was trained on nine folds (~11,097 rows) and tested on the remaining (~1,233 rows). This approach helps minimise bias and variance, ensuring that all observations are used both in training and evaluation.

Tuning Method: No complex tuning was required beyond standard preprocessing, as Logistic regression has relatively few hyperparameters. However, regularization was applied as it has in-built options such as L1 (Lasso) which was useful for penalising features that did not improve model accuracy.

Step 1: Baseline Model

- Includes all features from the dataset
- Ideal for benchmarking performance
- This model was refined using various techniques to create a parsimonious model that included meaningful predictors and lowered AIC.
- **AIC: 9,940**

Step 2:Tuning Parameters and Feature Selection using Lasso Regression (via glmnet)

- Purpose: penalise complexity and automatically selects features by shrinking coefficients to zero.
- Best tuning Parameters: Lambda (regularization strength) = **Seq(0.0001, 0.1, length.out = 10)** and Alpha = 1 (lasso penalty).
- Benefits: prevents overfitting and handles multicollinearity.

Step 3:Feature Selection using STEPWISE SELECTION (via AIC)

- Direction: Both forward and backward
- Criteria: Akaike Information Criterion (AIC)
- balanced models fit and complexity
- Lower AIC indicates better trade-off
- Benefits: selects parsimonious model with reasonable explanatory power.
- **AIC: 9,915**

Step 4: P-value Based Elimination

- Custom iterative function repeatedly removes variables with P-values > 0.05 (statistically insignificant)
- Stops when all the remaining predictors are statistically significant
- Benefit: ensures only meaningful predictors are kept for interpretability.
- Final model performance was compared with random forest and XGBoost model.
- **AIC: 9,921**
- Although there is a slight increase in AIC compared to stepwise model, this is outweighed by the model's reduced complexity and interpretability making it more practical for business use/decision-making.

Benefits of using all three approaches (tuning & feature selection):

- Robustness (Lasso)
- Simplicity and interpretability (StepAIC)
- Statistical validity (P-value filtering)

RANDOM FOREST MODEL



Overview: Random Forest is an ensemble learning algorithm that forms multiple decision trees and combines outputs (aggregating their results) to improve predictive accuracy and stability. It was selected for its robustness to overfitting, ability to model non-linear relationships and effectiveness in capturing complex interactions.

Training Method: Model training was performed using 10-fold cross-validation to ensure reliable performance across different data partitions.

Tuning Method:

- A grid search procedure was used to identify optimal hyperparameters, such as number of trees, tree depth, and minimum samples per split to maximise performance while reducing overfitting.

Table 1: Tuning Grid

| Parameters | Value |
|---|----------------------|
| mytry (number of variables tried at each split) | 2, 4, 6 |
| nodesize (minimum size of terminal nodes) | 5, 10 |
| Splitrule | gini (held constant) |

Table 2: Best Model Parameters

| Parameters | Best Tune Value |
|---|----------------------|
| mytry (number of variables tried at each split) | 6 |
| nodesize (minimum size of terminal nodes) | 5 |
| Splitrule | gini (held constant) |

X G B O O S T M O D E L



Overview: The XGBoost model is a powerful and efficient gradient boosting algorithm designed for speed and performance. Selected due to its reputation for delivering strong results in many classification tasks by combining gradient boosting with advanced regularization techniques to help overfitting.

Training Method: Model training was performed using 10-folds cross validation to ensure robust evaluation and generalisation across different subsets of data.

Tuning Method:

- The grid search approach was employed to systematically explore combinations of hyperparameters such as learning rate, max depth and subsample ratio. This process identified optimal settings that maximise model performance, balancing bias and variance.

Table 3: Tuning Grid

| Parameters | Tune |
|---------------------|--------|
| nrounds | 100 |
| max_depth | 4 to 6 |
| eta (learning rate) | 0.1 |
| gamma | 0 |
| colsample_bytree | 0.8 |
| subsample | 0.8 |
| min_child_weight | 1 |

Table 4: Best Model Parameters

| Parameters | Best Tune Value |
|---------------------|-----------------|
| nrounds | 96 |
| max_depth | 6 |
| eta (learning rate) | 0.1 |
| gamma | 0 |
| colsample_bytree | 0.8 |
| subsample | 0.8 |
| min_child_weight | 1 |



PR CURVE COMPARISON

Precision-Recall AUC as a key metric:

- Although the dataset used for modelling has been balanced, the original data was highly imbalanced, with far fewer positive cases (buyers) compared to non-buyers.
- Because of this initial imbalance, traditional accuracy metrics could still be misleading, as models might struggle to correctly identify the minority class.
- The Precision- Recall AUC provides a comprehensive summary of this balance, making it the most meaningful metric for assessing model performance in this context.
- To ensure a thorough evaluation additional performance metrics such as ROC AUC, F1 score and confusion matrix among others were used to analyse prediction errors.

Logistic Regression:

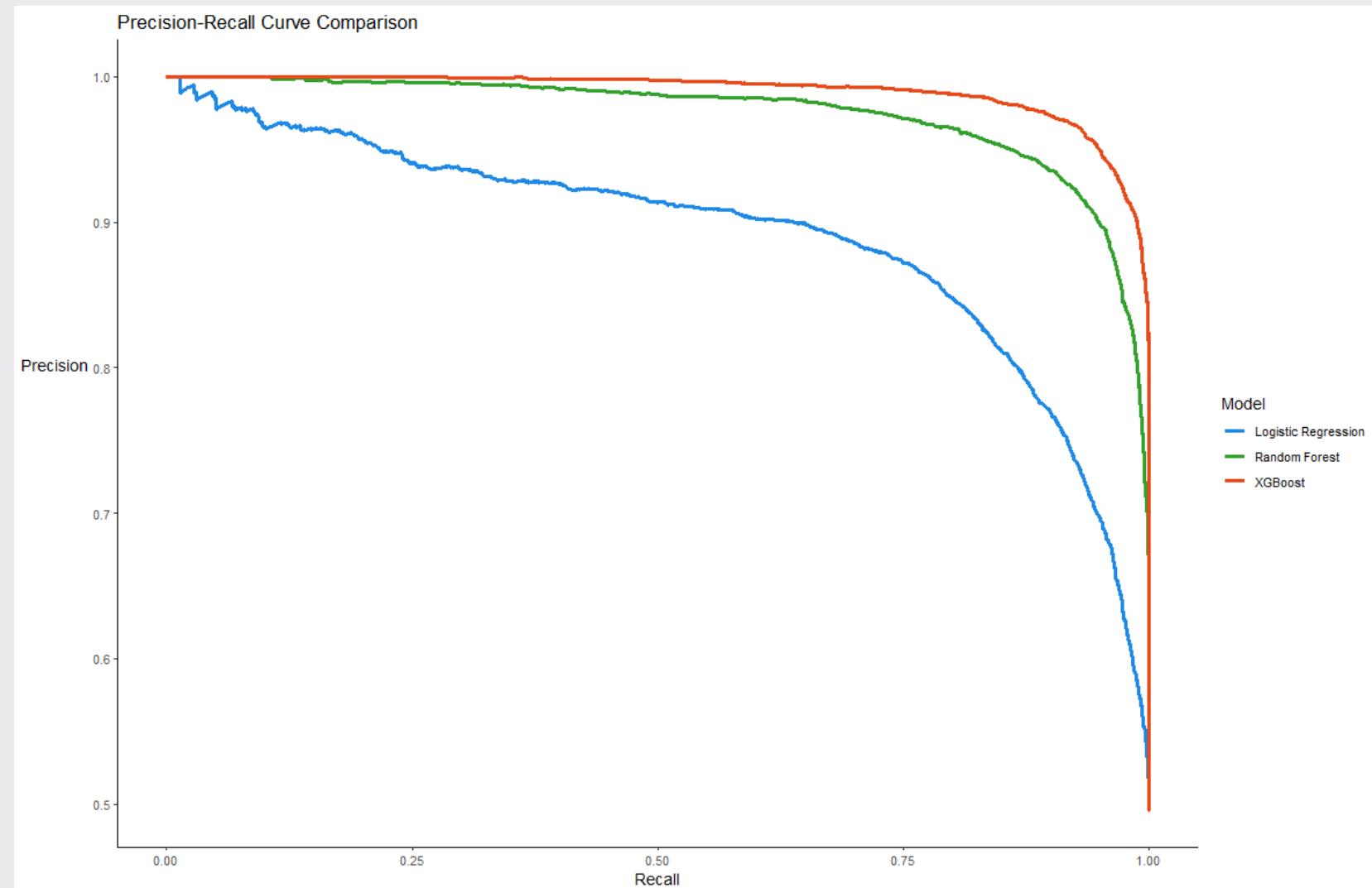
- Achieved a PR AUC of **0.89**, which indicates that the model is the least effective at distinguishing positive instances compared to the tree-based models. This suggests that it may struggle with capturing complex patterns in the data.

Random Forest:

- Improved performance with a PR AUC of **0.97**, demonstrating strong ability to identify positives while maintaining precision.

XGBoost:

- Best performing model with a PR AUC of **0.99**, showing excellent balance between precision and recall and capturing intricate relationships in the dataset.



PERFORMANCE METRICS



Logistic Regression Results:

- The logistic regression model was the lowest performing model across all classification metrics.
- The model produced an accuracy score of **0.83**, indicating that 83% of predictions were correct overall.
- In addition, the model achieved a recall of **0.81**, indicating that it correctly identified 82% of actual buyers.
- The precision score was **0.84**, suggesting that when the model predicted that users would buy, it was correct 84% of the time.
- The F1 score, which balances precision and recall was **0.83**.
- The ROC AUC was **0.91**, demonstrating strong ability to distinguish buyers and non-buyers, though this is not as high as the result for the tree-based models.

Random Forest Results:

- The random forest model showed improved performance by obtaining an accuracy score of **0.93**, indicating that 93% of predictions were correct overall.
- It had a recall of **0.92**, highlighting its strong ability to correctly identify actual buyers and a precision score of **0.93**, showing increased reliability in positive predictions.
- The F1 score of **0.93** demonstrates a well-balanced performance.
- The ROC AUC score of **0.98** indicates excellent separation between classes, reflecting the model's robustness in classifying buyer intent.

XGBoost Results:

- XGBoost delivered the strongest results by outperforming the other models as indicated by the key performance metrics.
- Accuracy was **0.95**, the highest among all models.
- Recall was the highest at **0.96** which indicates that the model correctly identified actual buyers (revenue-generating sessions) 96% of the time. This suggests that the model is highly effective at capturing the positive class and minimising the false negatives (**4%**) which is critical when the goal is to not miss potential buyers. Thus, this model is suitable as the priority is to maximise detection of actual buyers.
- The precision score was **0.94** which indicates that the model correctly predicted buyers as real buyers 94% of the time, which is an excellent score considering that it mostly avoids false positives.
- The F1 score was **0.95** which indicates a strong balance between precision and recall. This means it was both accurate in identifying actual buyers and effective at minimising false positives.
- The model achieved the same ROC AUC as the random forest model, indicating near-perfect ability to distinguish classes.
- These results are in line with the findings from Abdullah-All-Tanvir et al. (2023) and Deniz & Bülbül (2024). This consistency reinforces the effectiveness of ensemble models in handling behavioural prediction tasks in e-commerce contexts.

Comparison of Model Performance Metrics

| Model | Accuracy | Recall | Precision | F1 Score | ROC AUC | PR AUC |
|---------------------|----------|--------|-----------|----------|---------|--------|
| Logistic Regression | 0.831 | 0.813 | 0.841 | 0.827 | 0.902 | 0.891 |
| Random Forest | 0.924 | 0.920 | 0.925 | 0.923 | 0.977 | 0.974 |
| XGBoost | 0.949 | 0.956 | 0.942 | 0.949 | 0.977 | 0.990 |

STRENGTHS AND LIMITATIONS



| Model | Strengths | Limitations |
|---------------------|---|---|
| Logistic Regression | <ul style="list-style-type: none">Highly interpretable (coefficients show feature effects).Fast to train.Works well with linearly separable data. | <ul style="list-style-type: none">Lower predictive performance.Assumes linear relationships.May underperform with complex patterns. |
| Random Forest | <ul style="list-style-type: none">Handles non-linear relationships.Robust to overfitting.Good with imbalanced datasets (if tuned). | <ul style="list-style-type: none">Less interpretable than Logistic Regression.Slower to train with many trees.Performance not as high as XGBoost. |
| XGBoost | <ul style="list-style-type: none">High predictive accuracy.Handles complex feature interactions and non-linear relationships.Includes regularization. | <ul style="list-style-type: none">Complex to tune.Less interpretable (“black-box”).Can overfit if not managed. |

FEATURE IMPORTANCE



Importance Measurement by Model Type

Logistic Regression: Absolute Coefficient

- Magnitude of coefficient.
- Larger the absolute coefficient means that a feature has more influence on the likelihood of the outcome.

Random Forest: Permutation

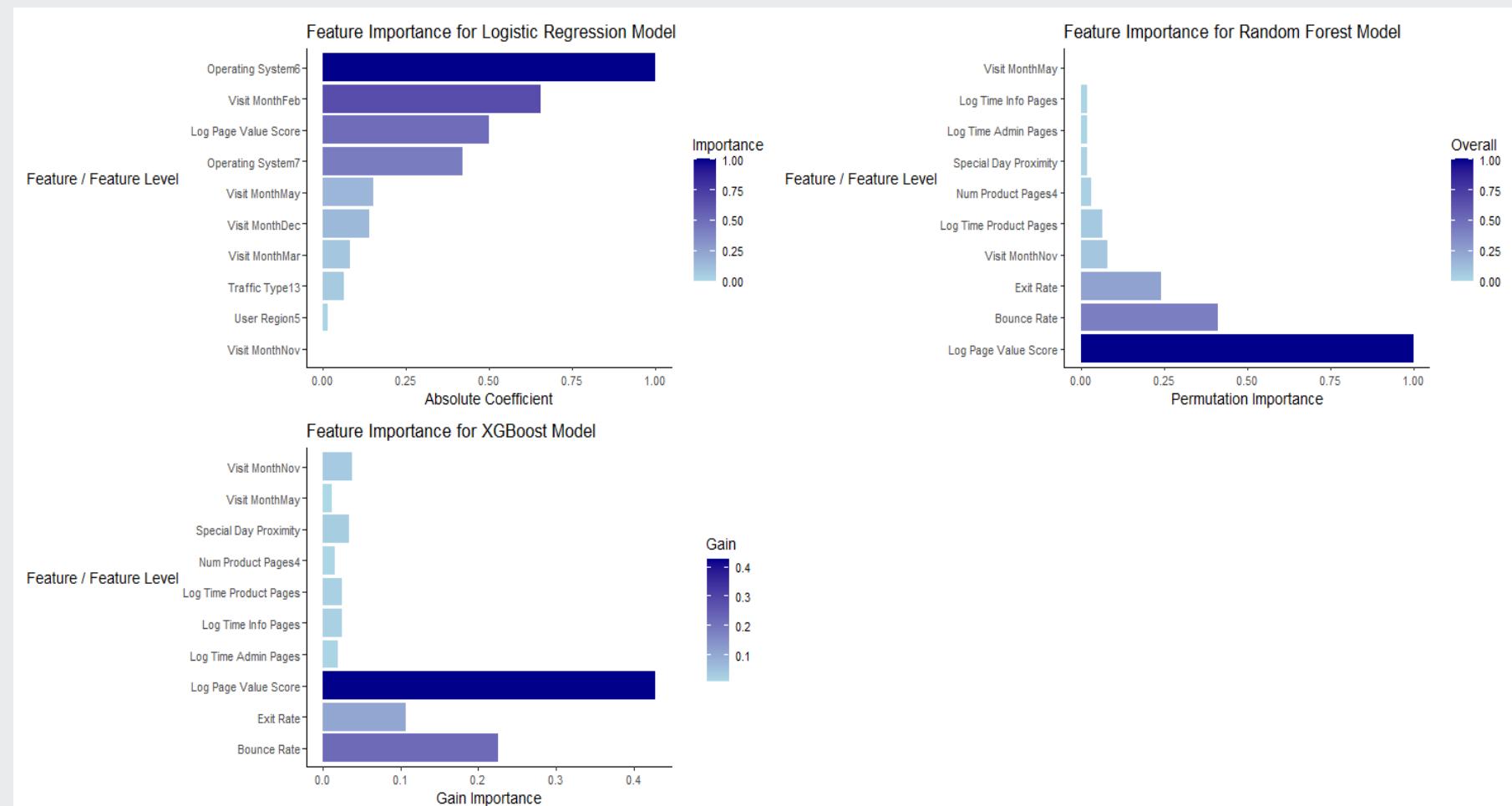
- Randomly shuffles one feature at a time and sees how much the model performance drops
- If the performance drops significantly, the feature was important.

XGBoost: Gain

- It represents the improvement in performance accuracy (reducing loss) brought by a feature each time it's used to split.
- Higher gain implies that a feature is informative.

Results:

- Among all features, the XGBoost model determined that page value score was the most important predictor of purchase likelihood, followed by bounce rate, exit rate and visit month November.
- The results for both tree-models (XGBoost and Random Forest) were very similar. Reflecting their ability to effectively capture the complex and non-linear relationships.
- However, the logistic regression model had very different results, with operating system 6 the most important predictor followed by visit Month Feb and page value score.

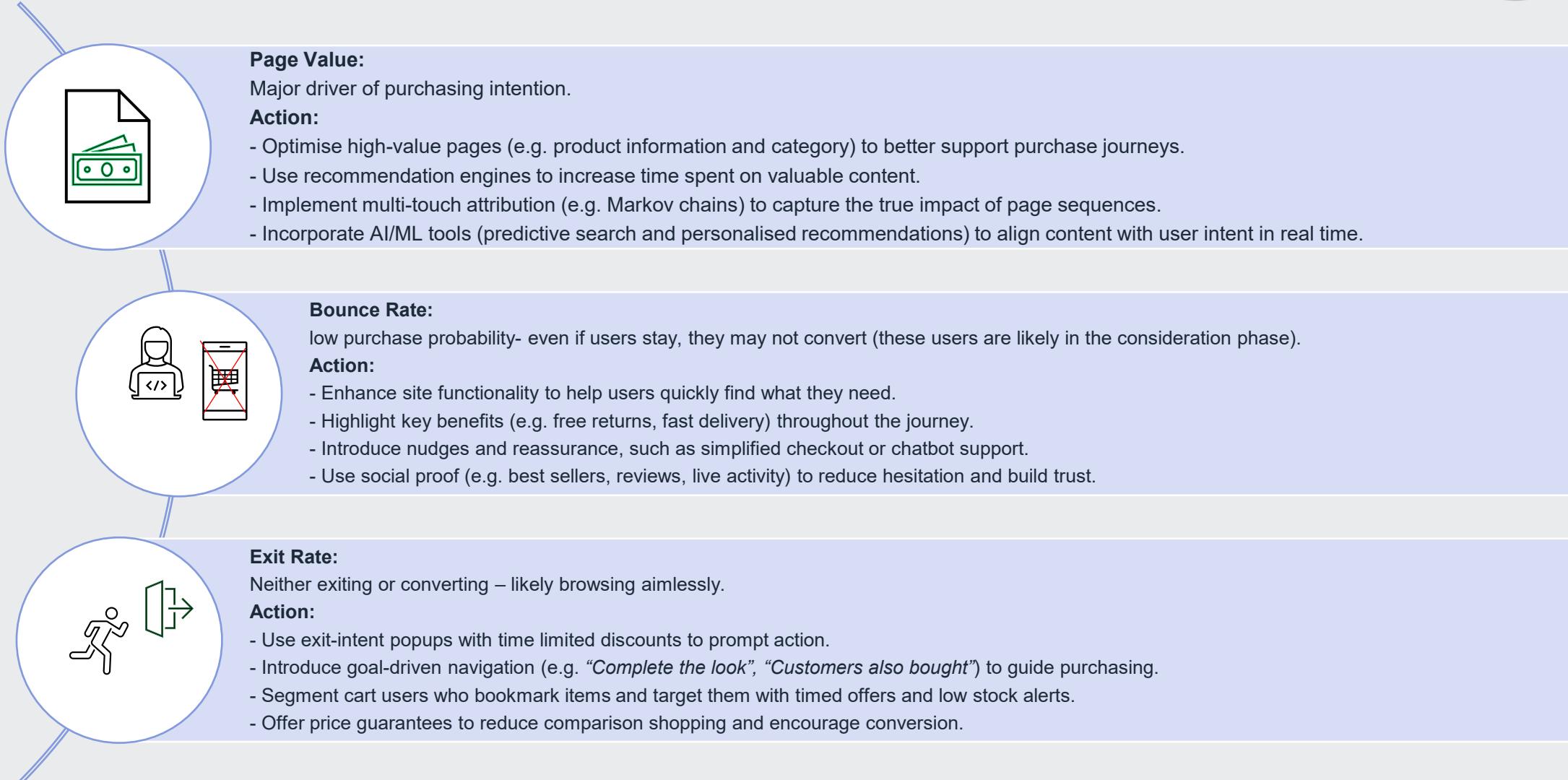


CONTEXTUALISING MODEL EXPLANATIONS



| Feature | Model Statistical Insight (SHAP Value) | Comparison |
|------------------|--|---|
| Page Value (log) | <ul style="list-style-type: none"> The strongest predictor of purchasing intent obtained a mean Shap value of approximately 0.2, indicating a moderate positive contribution to the model's prediction. In this case, higher values of the feature increase the likelihood of conversion. This suggests that greater user engagement with high-value pages is strongly associated with purchasing behaviour, reinforcing the importance of content-rich interactions in driving conversions. | <ul style="list-style-type: none"> The results align with prior research that emphasises the importance of user engagement and content-rich page interactions in driving purchasing intent. For example, Close and Kukar-Kinney (2009) identified motivations such as information search, entertainment and organisation as key drivers of online cart use and purchase behaviour – suggesting that engagement extends beyond immediate transactional intent. Similarly, Markov chain-based attribution models recognise the role of product and category pages in shaping the user's journey, even if they are not the final touchpoint before conversion. This finding that higher page value scores yield positive SHAP values (~0.2) supports this view, indicating that value-rich, informative pages contribute meaningfully to purchase intent, even if they are not directly associated with the final purchase click. Thus, the SHAP analysis reinforces the idea that user interactions across multiple high-value pages cumulatively influence conversion, highlighting the importance of moving beyond last-click attribution to understand purchasing behaviour holistically. |
| Bounce Rate | <ul style="list-style-type: none"> Interestingly, the obtained a mean SHAP value indicates a low to moderate influence (approximately -0.1 or less). The negative sign suggests that this feature level is associated with a lower probability of purchase. While the effect is not strong, it does imply that users represented by this feature level are less likely to convert. This may reflect browsing/exploratory behaviour without strong purchasing intent – users are engaging with the site but not exhibiting actions that significantly drive purchase predictions. | <ul style="list-style-type: none"> The model insights resonate with findings from Cialdini (2001), who noted that users initially motivated to browse can be persuaded to buy later through attractive incentives. This suggests that while the model predicts low conversion likelihood for these browser users, they represent a segment with potential for conversion if targeted effectively with offers or promotions. Thus, the combination of SHAP-based model interpretation and behavioural research highlights an opportunity to nurture browsers into buyers with well-timed incentives. |
| Exit Rate | <ul style="list-style-type: none"> Similarly, the mean SHAP value indicates a low to moderate influence (approximately -0.1 or less). The negative sign indicates a small downward contribution to the model's purchase likelihood. This suggests that users with this exit rate are less likely to purchase, although the effects are minimal. These results may imply that users are not exiting the site immediately, but also not taking actions associated with conversion – reflecting low purchase intent despite some level of engagement. | <ul style="list-style-type: none"> The results for the mean SHAP value for exit rate align with prior findings. For example, Close and Kukar-Kinney (2010) observed that many users use online carts as wish lists or for comparison, not for immediate purchases. Similarly, hedonic browsers (those engaging with for entertainment) often stay longer without goal-directed actions. As such, simply remaining on the site does not necessarily reflect high purchase intent, echoing what the SHAP value suggests. Moreover, Shafir et al. (1993) found that visible, time-limited discounts can successfully convert general browsers into buyers. McConnell et al. (2000) also noted price guarantees can discourage comparison shopping and drive users to complete purchases. Together, these findings support the interpretation that prolonged engagement without conversion signals low intent, but with the right incentives (e.g. limited time offers or price reassurances) this segment can still be influenced toward conversion. |

RECOMMENDATIONS AND NEXT STEPS





UNIVERSITY OF
SURREY

THANK YOU !!

