BT5151 - ADVANCED ANALYTICS AND MACHINE LEARNING (ASSIGNMENT 2)

INCREASING PURCHASE CONVERSION RATE

Group 12

Georgius Gary Gunawan (A0113028W)
Lam Yan Tung (A0086956J)
Patricia Tay Li-Min (A0057228A)
Susan Koruthu (A0231905L)
Widya Gani Salim (A0231857Y)

CONTEXT & PROBLEM STATEMENT

3 **Pressing issues faced by retail companies**

- Drastic shift from retail shopping to e-commerce
- Conversion rates have not increased proportionally
- Rising needs to analyze online purchasing behaviour

46%*

Growth of e-commerce proportion as a total retail sales worldwide

12.20% in 2018 → 17.80% in 2020
(absolute increase of 5.60%)

-27%#

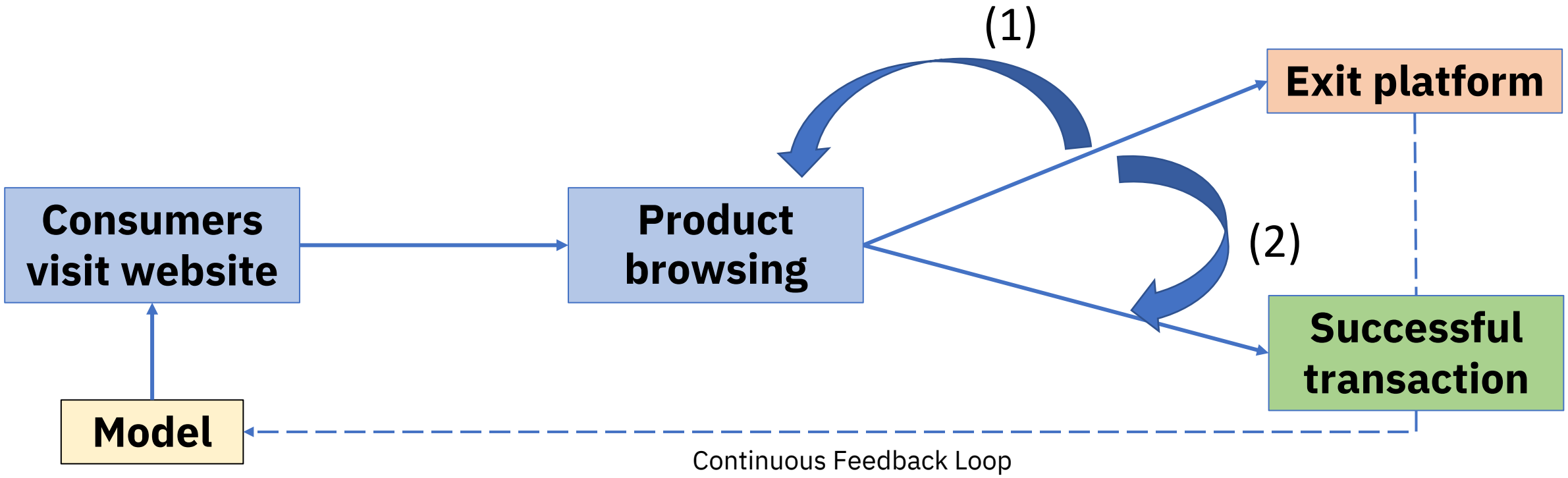
Conversion rate of online shoppers worldwide

2.97% in 2018 → 2.17% in 2020
(absolute decrease of 0.80%)

Key Objective: Predicting online user's purchasing intention, whether they will convert.

*Source: https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/
#Source https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/

INTEGRATION TO E-COMMERCE



By predicting user's purchasing intention, e-commerce operations will improve by:

(1) Preventing Churn: Introduce marketing campaigns & advertisements to high value pages
(2) Increasing Purchasing Conversion: Improve or scrap low value or leaky webpage.

Prediction model seamlessly integrate with existing operations.

AVAILABLE DATASETS

12,330

Visitor Sessions

Each session represents a different user within 1-year period
(to avoid user, time and campaign bias)

ATTRIBUTES

Numerical		Categorical	
Administrative Duration	Administrative	Operating System	Visitor Type
Informational Duration	Informational	Browser Type	Traffic Type
Product Related Duration	Product Related	Region	Revenue
Bounce Rate	Page Value	Weekend	
Exit Rate	Special Day	Month	

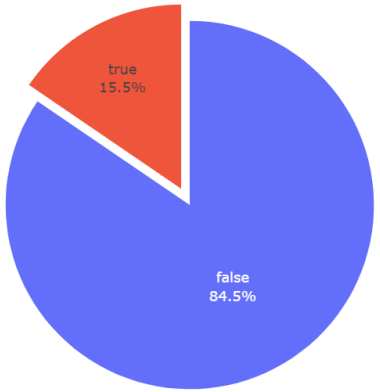
Time spent related user activity · Page related user activity · Date control variables

Individual characteristics/demographics · Target variable

EXPLORATORY DATA ANALYSIS

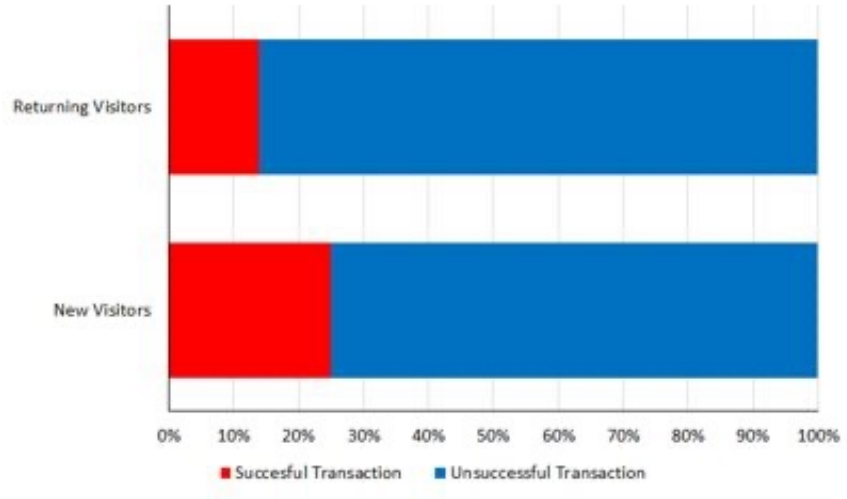
Imbalanced Dataset

Only 3 in 20 visits are successful transaction.

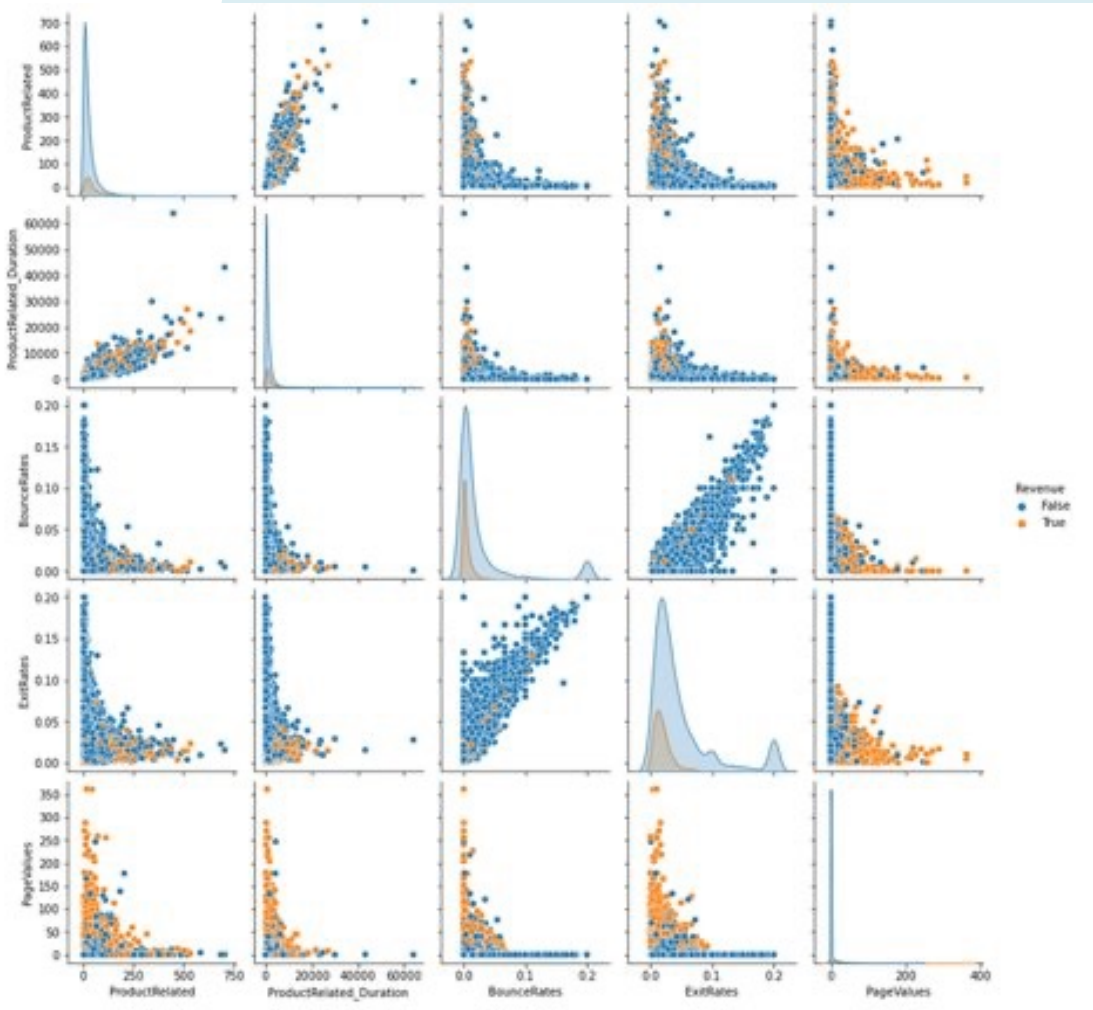


Propensity in Successful Transaction per Visitor Type

Returning visitors are half as likely to complete their transactions, compared to new visitors



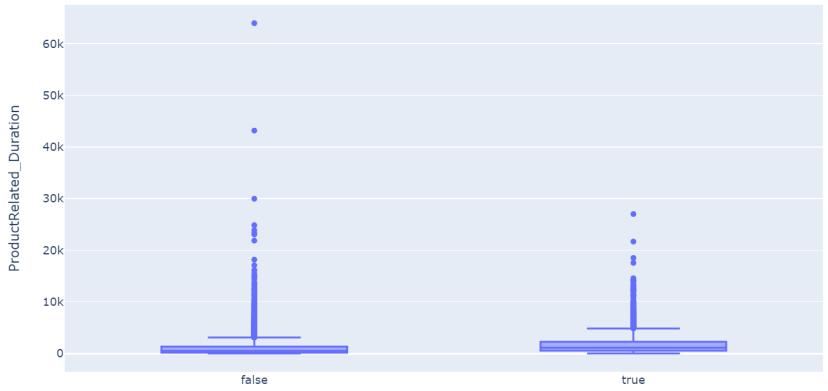
- **High Correlation among some features:** (e.g: BounceRates and ExitRates)
- **Correlation with Labels:** Some features (e.g. PageValues, ExitRates) show clear distinction in behaviors between successful and unsuccessful transactions.



EXPLORATORY DATA ANALYSIS

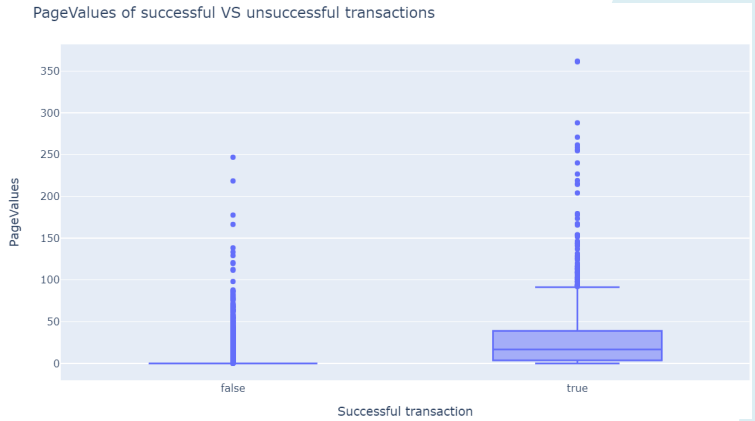
Outliers in ProductRelated_Duration

Users leaving opened product page idle



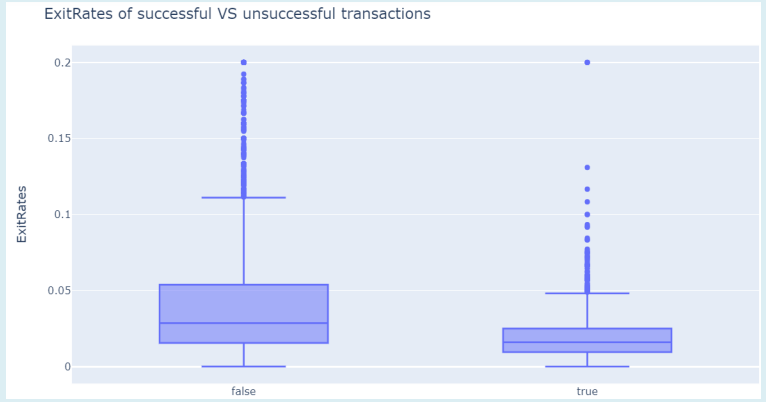
High PageValues for Positive Cases

Successful transactions have **16x** mean PageValues as compared to unsuccessful.



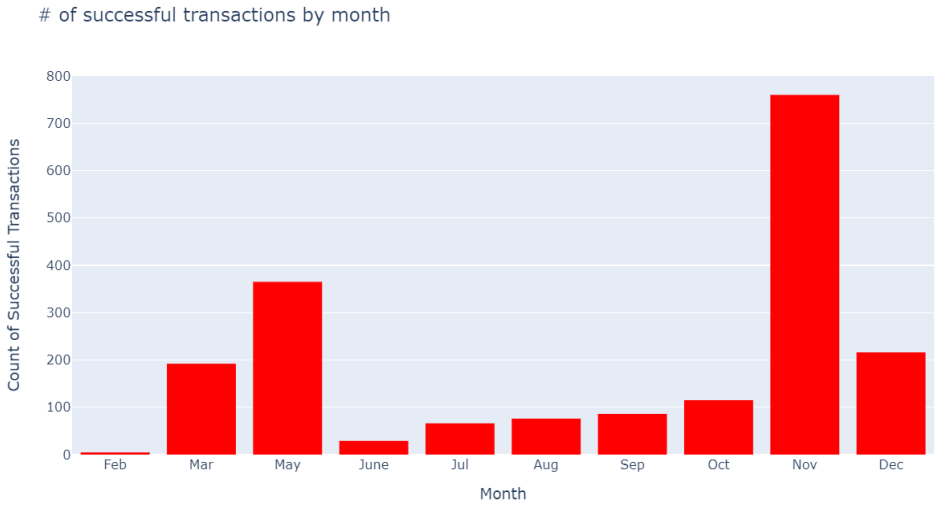
Low ExitRates for Positive Cases

On average, ExitRates for successful transactions is **3x lower** as compared to unsuccessful.



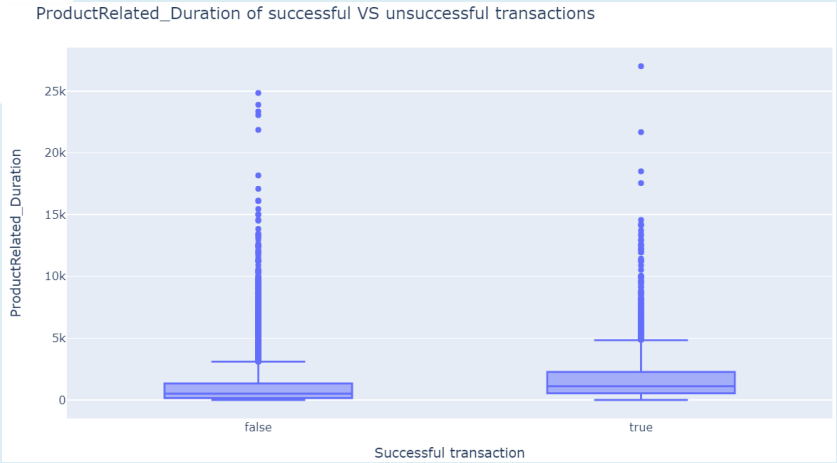
Month_Nov have the highest number of transactions

Amounting to more than 2x as compared to other months



High ProductRelated_Duration for Positive Cases

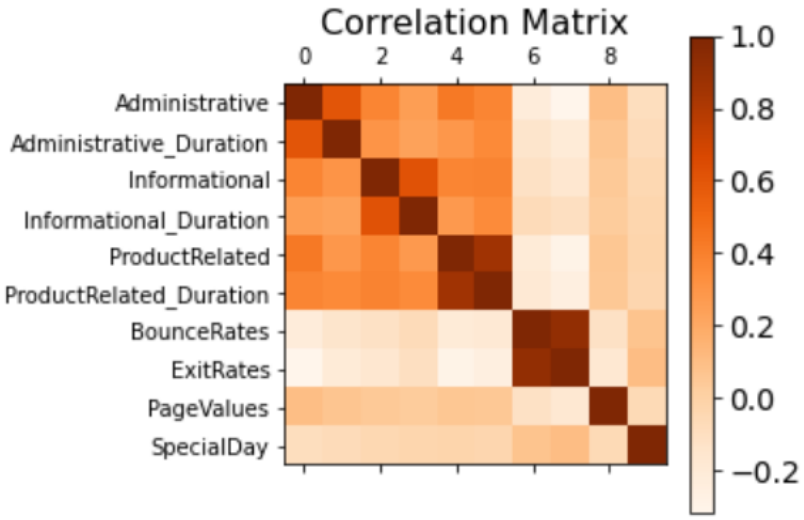
Mean of ProductRelated_Duration for successful transactions are **2x higher** as compared to unsuccessful.

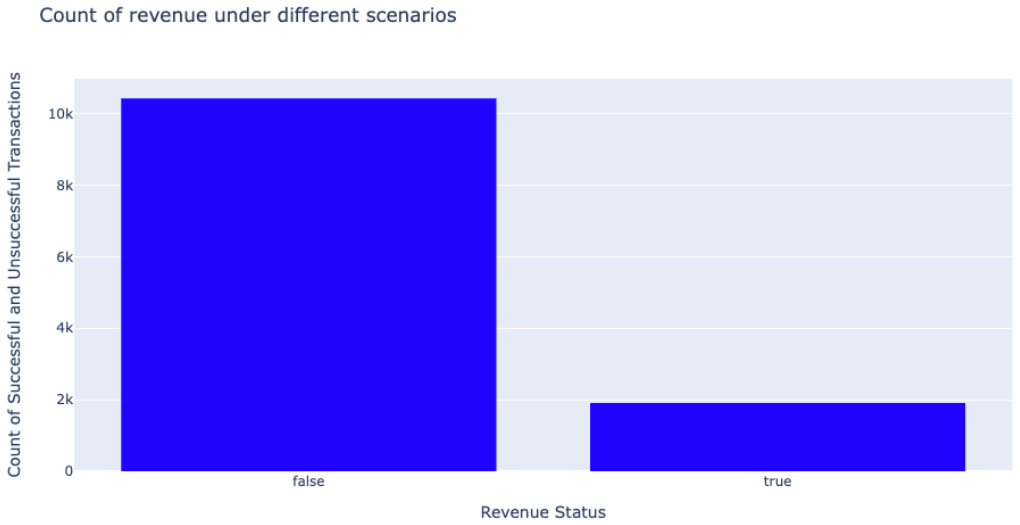


DATA PREPROCESSING

A few crucial steps were done to clean and transform the data in order to improve model fit and accuracy.

- **Removal of Outliers:** Outliers in all time-related activity variables (Administrative, Informational and Product Related Duration) were removed.
- **Categorical Features One-Hot Encoding**: Ensure categorical features can be inputted to selected models.
- **Z-score Standardization**: Rescaled some features to ensure that all features are of the same scale.
- **Resampling Imbalanced Dataset**: SMOTE technique was used to balance the original dataset.
- **Removal of Highly Correlated Features:** Based on heatmap to remove bias.
- **Test-Train Data Split:** Create train, validation and test sets to train and test the model more accurately.





FEATURE SELECTION

To boost performance and reduce computing cost, irrelevant features were eliminated.

18 } **66** } **9**

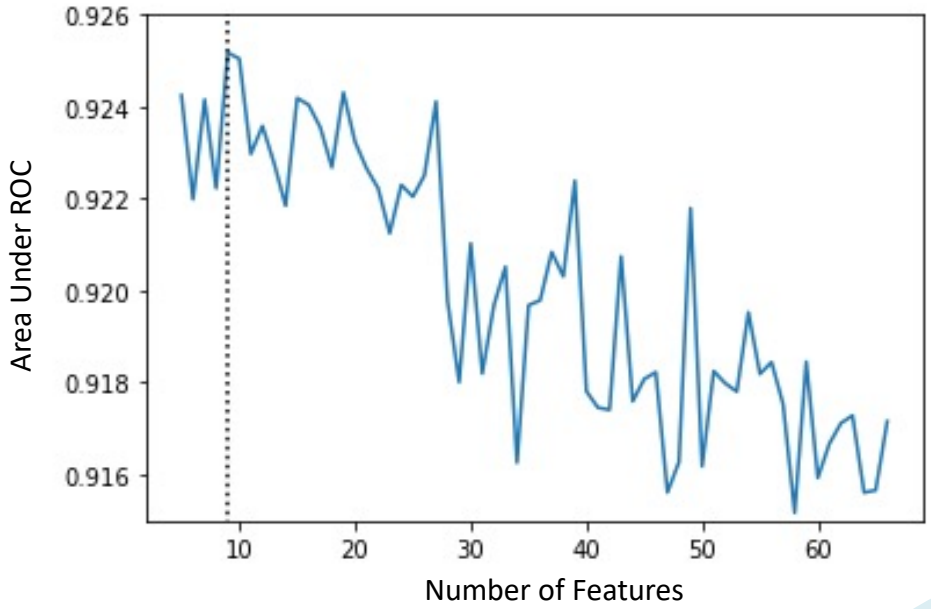
Original Features | **Engineered Features** | **Selected Features**

Recursive Feature Selection (RFE)
Removed least important features based on cross-validation performance on a selected model (Random Forest with max depth of 5).

RFE with Random Forest Max Depth = 5 was chosen

Selected Features: Administrative, Administrative_Duration, Informational_Duration, ProductRelated_Duration, ExitRates, PageValues, Month_Nov, VisitorType_New_Visitor, Month_May

No. of selected features vs AUC score



MODELS SELECTION & TRAINING

Explored 4 Classification & 1 Stacked Models:

- Logistics Regression (Baseline Model)
- Random Forest (Ensemble Bagging Method)
- XGBoost (Ensemble Boosting Method)
- Neural Network (Deep Learning)
- Stacking of the above models (Stacking Model)

Selected Performance Metrics:

- ROC-AUC Score – Immune to bias due to the size of test data
- F1-score – Robust against imbalanced dataset

Hyperparameter Tuning Methods:

- Bayes Search CV
- Grid Search CV

HYPERPARAMETERS TUNING

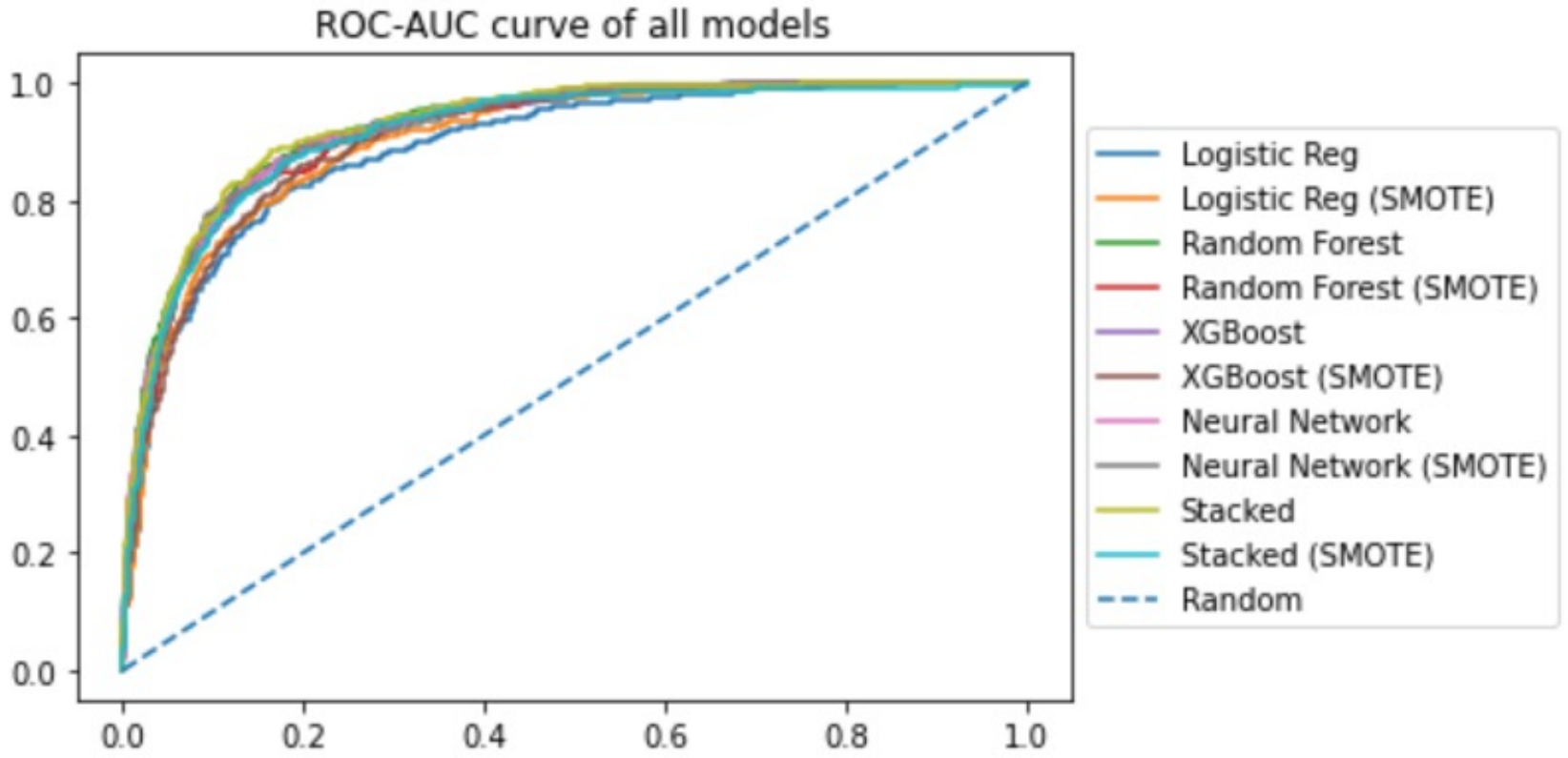
Model Name	Tuned Technique	Best Parameters	
Logistic Regression	Bayes Search CV	C = 0.1	Tol = 0.000350
Logistic Regression (with oversampling)		C = 99.25	Tol = 0.0153
Random Forest		Criterion = entropy	Min_samples_leaf = 5
		Max_depth = 8	Min_samples_split = 10
Random Forest (with oversampling)		Criterion = entropy	Min_samples_leaf = 2
		Max_depth = 24	Min_samples_split = 5
XGBoost		colsample_bylevel = 0.592	Max_depth = 9
		Colsample_bytree = 1.0	Min_child_weight = 5
		Gamma = 1e-09	N_estimators = 76
		Learning_rate = 0.01	Subsample = 0.788
XGBoost (with oversampling)		colsample_bylevel = 1.0	Max_depth = 15
		Colsample_bytree = 0.637	Min_child_weight = 0
		Gamma = 1.36	N_estimators = 100
		Learning_rate = 0.0935	Subsample = 0.82
Neural Network	Grid Search CV	batch_size = 50	Learning_rate_init = 0.01
		Hidden_layer_sizes = (6,6,6)	Tol = 0.0001
Neural Network (with oversampling)		batch_size = 10	Learning_rate_init = 0.01
		Hidden_layer_sizes = (12,8,1)	Tol = 0.0001

EVALUATION & METRICS

	Logistic Regression	Logistic Regression (with SMOTE)	Random Forest	Random Forest (with SMOTE)	XGBoost	XGBoost (with SMOTE)	Neural Network	Neural Network (with SMOTE)	Stacking	Stacking (with SMOTE)
Accuracy	0.87	0.85	0.90	0.88	0.89	0.88	0.89	0.89	0.89	0.88
Precision	0.74	0.54	0.75	0.63	0.73	0.65	0.76	0.76	0.74	0.64
Recall	0.36	0.76	0.59	0.73	0.59	0.69	0.53	0.53	0.56	0.70
F1 Score	0.49	0.63	0.66	0.68	0.65	0.67	0.62	0.62	0.64	0.67
AUC	0.890	0.899	0.927	0.917	0.925	0.916	0.923	0.923	0.928	0.912

- Better F1-score when over-sampling is applied on all models, except Neural Network.
- All models have better AUC score compared to baseline model.
- Performance between models with SMOTE are comparable, but Random Forest is the most balanced between its F1 and AUC scores.

EVALUATION & METRICS



- Performance of all models are comparable, with a difference of +/- 4% in AUC.
- Stacked model has the best AUC as discussed. However, we are selecting Random Forest (with SMOTE) as it gives the most balanced score and is comparable to the stacked model.

RESULT & INTERPRETATION

Model (with SMOTE)	Logistic Regression	Random Forest	XGBoost
Top 5 features	PageValues	PageValues	PageValues
	Month_Nov	ExitRates	Month_Nov
	ExitRates	ProductRelated_Duration	ProductRelated_Duration
	VisitorType_New_Visitor	Month_Nov	Administrative
	Month_May	Administrative_Duration	ProductRelated_Duration

Results are in line with findings in EDA:

- All models indicate PageValues as the top important feature to predict purchasing behavior
- ExitRates, month_Nov, and ProductRelated_Duration also come up at high rankings in all models.

RECOMMENDATION & NEXT STEPS

Confusion matrix is computed based on Random Forest (with SMOTE) model.

Cost-benefit values is calculated based on average transaction value and average cost per conversion*.

Confusion Matrix	**True Positive**	**True Negative**
Predicted Positive	305	180
Predicted Negative	117	1864

Cost-benefit Values	**True Positive**	**True Negative**
Predicted Positive	$50	- $25
Predicted Negative	$0	$0

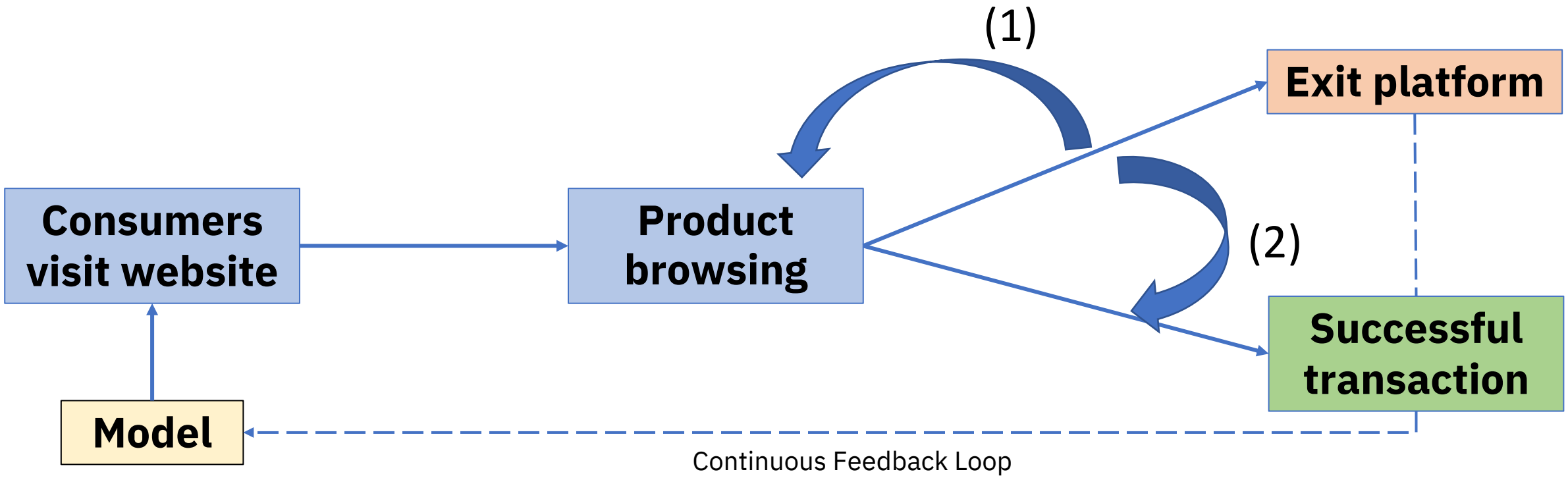
Baseline Expected Revenue (Random Case): ((15.5/100)*$50) - ((84.5/100)*$(25)) = -$13.38 / user

Expected Revenue with Model: ((305/2466)*$50) - ((180/2466)*$(25) = $4.36 / user

$17.74 INCREASE IN REVENUE PER USER!

*Data is retrieved from https://guidingmetrics.com/content/ecommerce-industry-most-critical-metrics-kpis/

RECOMMENDATION & NEXT STEPS



In trial phase, prediction model can run in the backend. This will help the e-commerce platform to **prevent churn** and **increase conversion** by detecting potential purchase loss early **without affecting existing operations**.

Continuous feedback loop as more consumers visit the website will improve model performance.

THANK YOU!
ANY QUESTIONS?

Group 12
Georgius Gary Gunawan (A0113028W)
Lam Yan Tung (A0086956J)
Patricia Tay Li-Min (A0057228A)
Susan Koruthu (A0231905L)
Widya Gani Salim (A0231857Y)