ANALYSIS AND PREDICTION OF ONLINE SHOPPERS' PURCHASE INTENTION USING VARIOUS ALGORITHMS

Mentor –

Mr. Suvajit Mukhopadhyay

PRESENTED BY –

Amit Sharma

Amitesh Bajpai

Kushal Maheswari

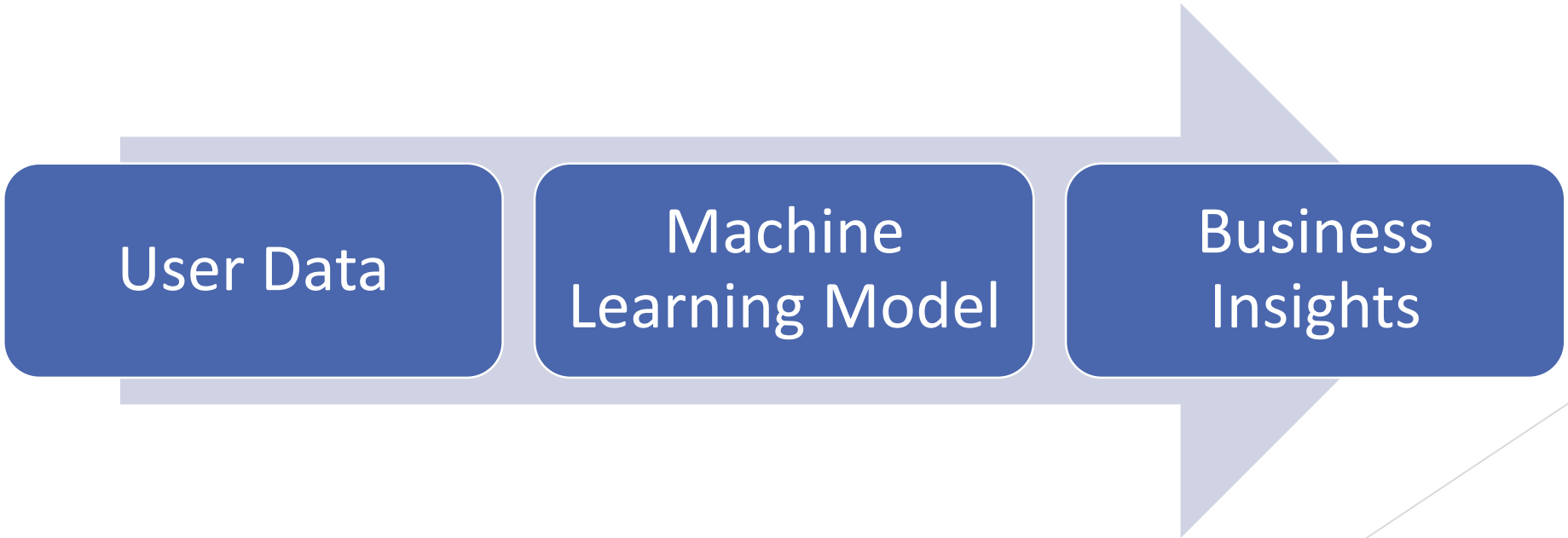
Neeraj Singh

Sahil Sachdeva

TABLE OF CONTENTS

Introduction

- E-commerce, the activity of buying and selling products online, is one of the many fields revolutionized by data science.
- One of the essential goals for e-commerce companies is to increase purchase conversion rates, i.e. the percentage of website visitors who complete the purchase at online stores.
- To achieve this goal, e-commerce companies as well as researchers in academia have devoted efforts in analyzing and modelling the behaviors of webpage users.
- Especially in recent year, there has been a trend in research to use machine learning methods to predict the behavior of users.



Scope

Phase 1 → Phase 2

Problem Statement Definition

- Define tentative objectives
- Define Purpose
- Expectation and Application

Understanding the Data

- Preparation and Preprocessing
- Dataset completeness test
- EDA
- Feature Engineering

Insights & Recommendation

- Best machine learning model selection
- Discover hidden patterns through analysis
- Predict Conversion rate
- Future Scope and Real-World Application

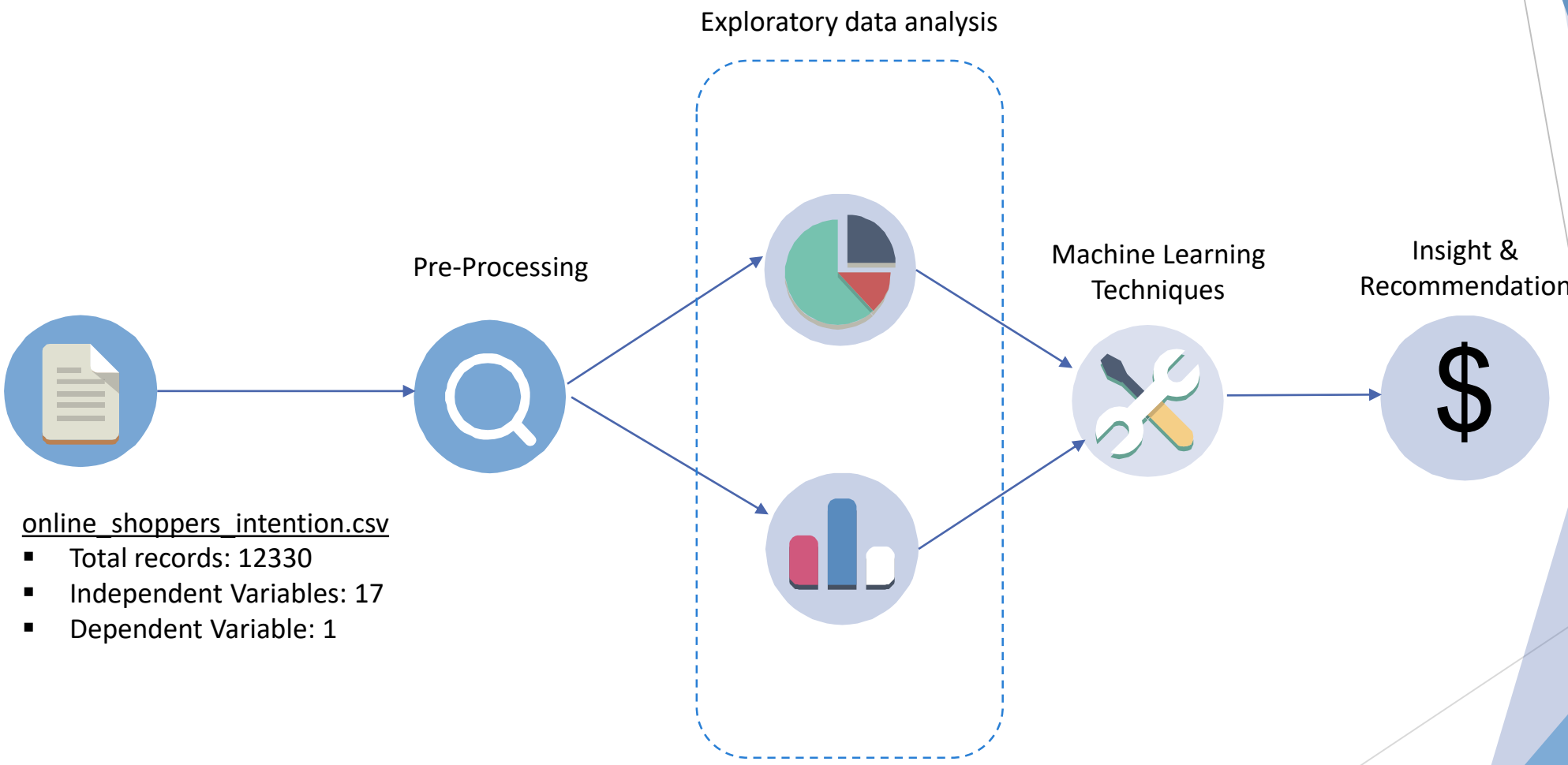
Objective

- ✓ Our overall objective is to analyze the given data and predict/put forward a reasonable action plan for the company which it can employ around its marketing strategies so that increase number of online visitors complete their purchase successfully.
- ✓ This will result in increased revenue generation and greater market share for the company.

STEPS

- Predict Shoppers' Intention to complete transaction.
- Classify user according to there behavior as revenue generating and non-revenue generating.

Overall Approach



online_shoppers_intention.csv

- Total records: 12330
- Independent Variables: 17
- Dependent Variable: 1

About The Dataset

Behavior

- Pages Visited (float64)
  1. Administrative
  2. Informational
  3. ProductRelated
- Page Duration (float64)
  1. Administrative_Duration
  2. Informational_Duration
  3. ProductRelated_Duration
- Visitor Type (object)
  1. New_visitor
  2. Returning_visitor
  3. Others

Google Analytics Metric

- BounceRate (float64)
- ExitRate (float64)
- PageValues (float64)

Campaign

- SpecialDay (float64)
- Month (object)
  JAN to DEC
- Weekend (bool)
  TRUE or FALSE

Geo

- Region (int64)
  1 to 9
- TrafficType (int64)
  1 to 20

Technology

- Browser (int64)
  1 to 13
- OperatingSystems (int64)
  1 to 8

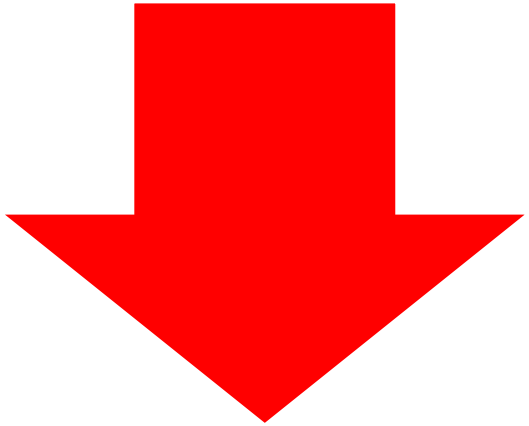
Target Variable – REVENUE *(bool)*

Data Source : *https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset*

Brief Overview of Data

- The dataset has users' web session details over one-year period.
- Out of 18, 8 features have 14 null values each. Null value treatment is to be done for these entries.

1908 users completed the transaction (15.5%)

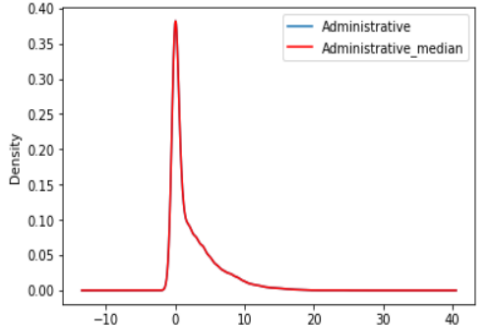


10422 users did not complete the transaction (84.5 %)

Data Processing

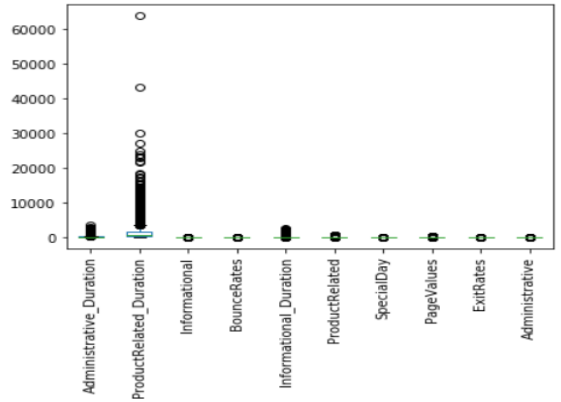
- **Missing Values** are replaced with Median. As there is no significant difference in Standard Deviation and Distribution. Since the data is skewed Median imputation was considered. *(Ref: Pg. 16 of report)*

```
Original Standard Deviation:  3.3227538813737088
Standard Deviation after median imputation:  3.3211633492176733
Standard Deviation after zero imputation:  3.32178410615674
Standard Deviation after random imputation:  3.3219763431542875
Standard Deviation after mean imputation:  3.320866795328869
```



- **Negative Values** are present in Administrative_Duration, Informational_Duration and ProductRelated_Duration columns. These columns represents the time spent on these pages, hence these value cannot be negative so they are replaced with zero. *(Ref: Pg. 18 of report)*

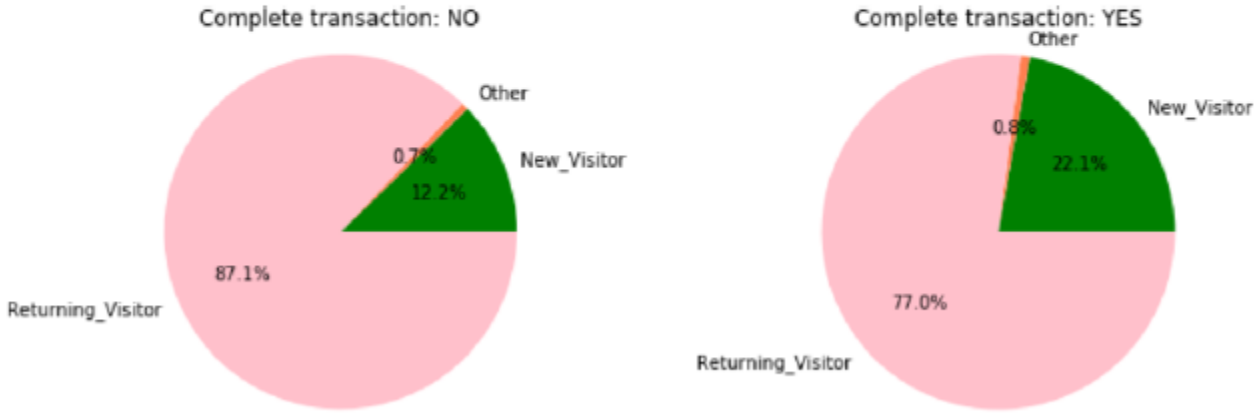
- **Outliers** are present in Administrative_duration, Informational_Duration and ProductRelated_Duration columns, Hence we decide to do top coding of Administrative_Duration at 2500, Informational_Duration at 2200 and ProductRelated_Duration at 20000. *(Ref: Pg. 20 of report)*



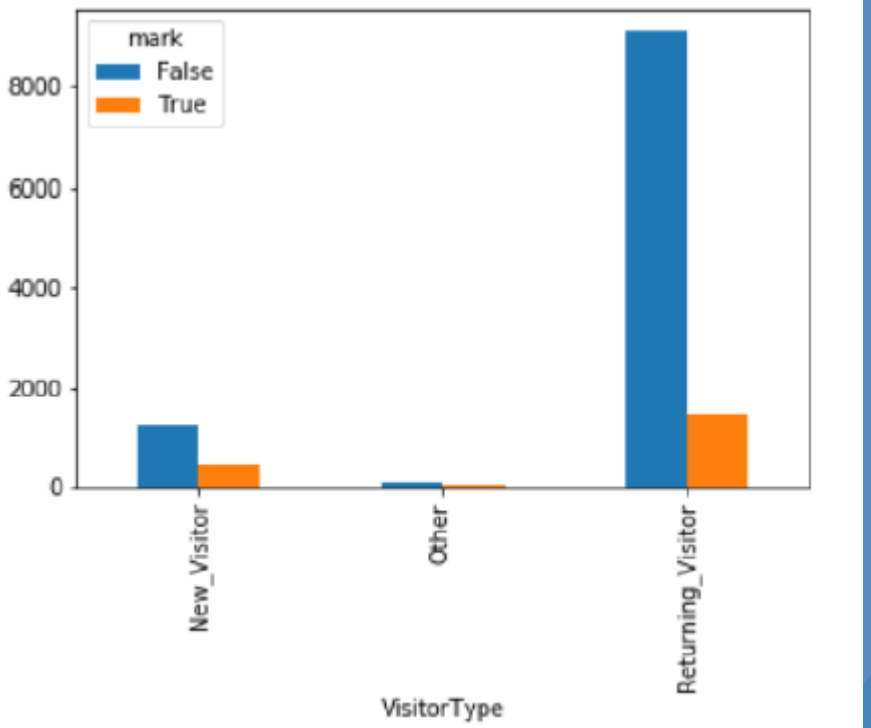
EDA : New Vs Returning Visitors

VisitorType	New_Visitor	Other	Returning_Visitor
Revenue			
False	1272	69	9081
True	422	16	1470

No of visitor types w.r.t. Revenue



Share of Visitor types w.r.t. Revenue



Insights

- Most of the customers visiting the website are Returning visitors, contributing to most number of purchases.
- About 1/4th of New_visitors made the purchase as compared to ~15 % of Returning visitors.
- There are more returning visitors among those who do not complete the purchase.

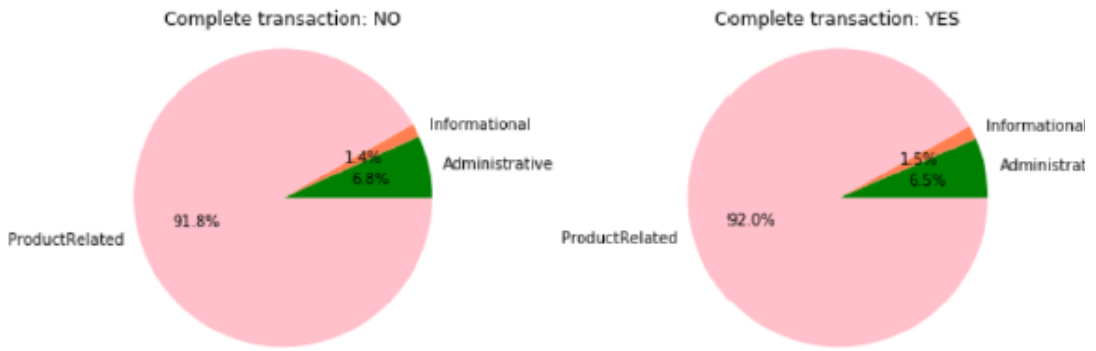
Insights - Page Visit and Time Spent

Revenue	Administrative	Informational	ProductRelated
False	2.120580	0.452248	28.748943
True	3.393606	0.786164	48.210168

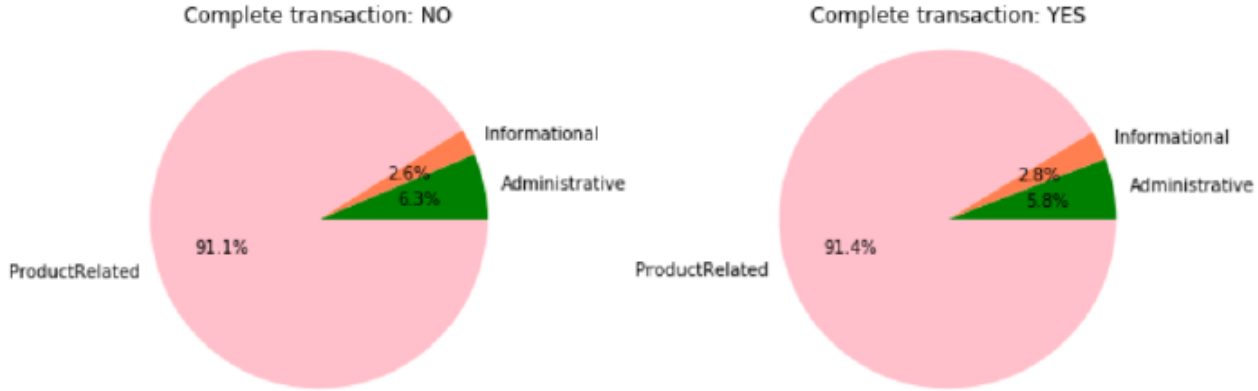
Avg visits on each page w.r.t. Revenue

Revenue	Administrative_Duration	Informational_Duration	ProductRelated_Duration
False	73.834208	30.270759	1071.347468
True	119.483244	57.611427	1876.209615

Avg duration spent on web page w.r.t. Revenue



Share of web page visits w.r.t. Revenue



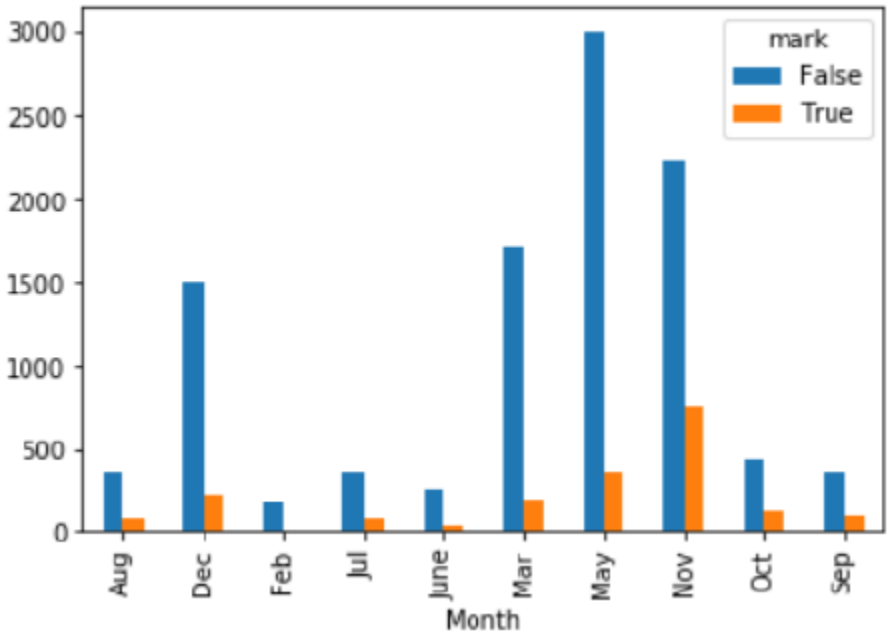
Share of duration spent on webpage w.r.t Revenue

Insights

- People visit product related pages more rather than Administrative or Informational which is obvious as user intends to purchase rather than just read admin and info pages on the website.
- On average, people who complete transactions visit more webpages, and spend more time on these webpages.

Insights: Impact of Month and Special Days

Monthly Conversion Rate



Most Active Months

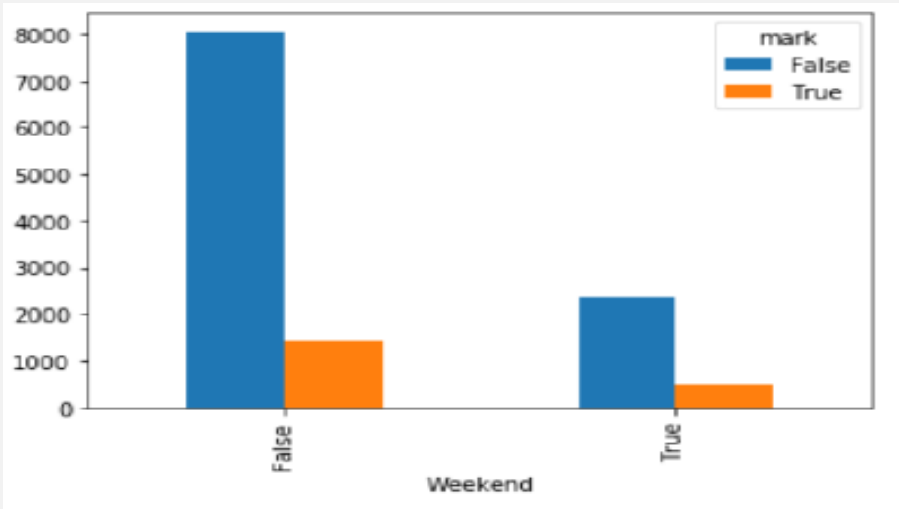
May | November | March | December

~81% Users (Combined)

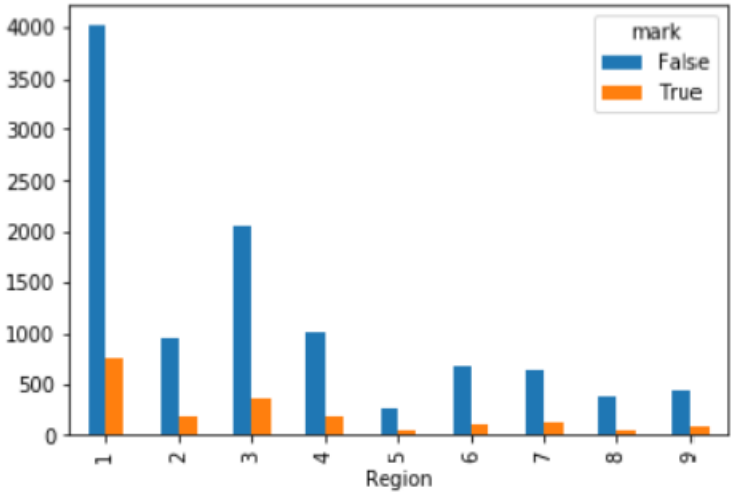
Special Days Impact?

~90% Interactions on Non-Special Days

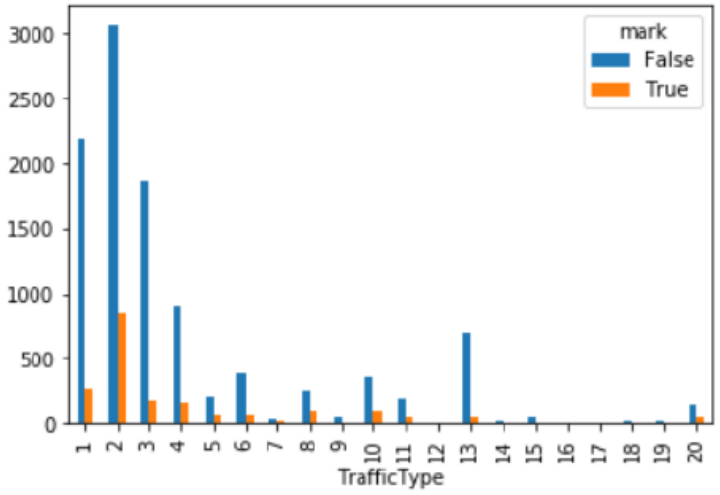
Online Shopping on Weekend?



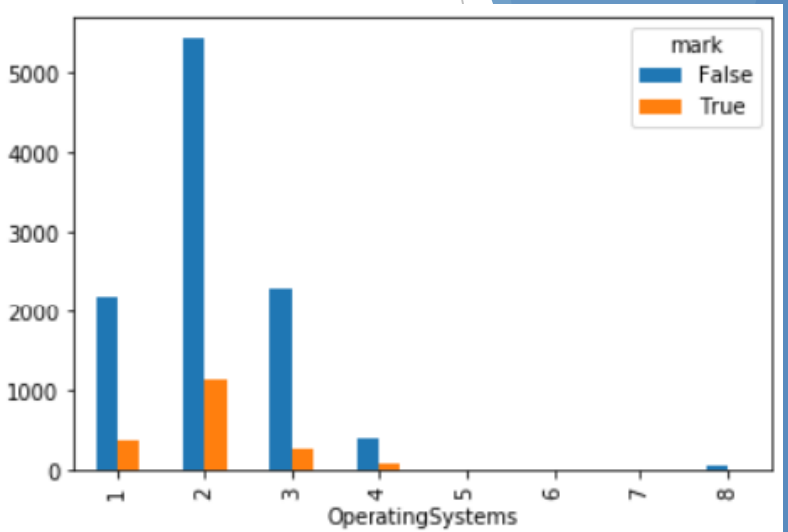
Insights: Impact of Region, Traffic Type, OS & Browser



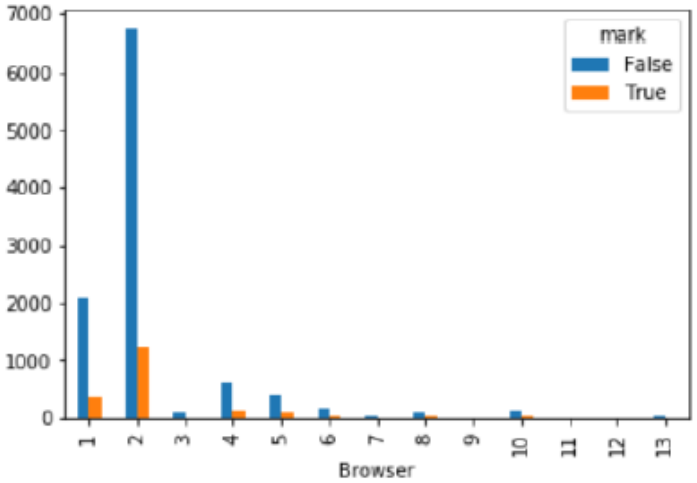
Most Populous Region



Most Traffic Generated Type



Most Used Operating System

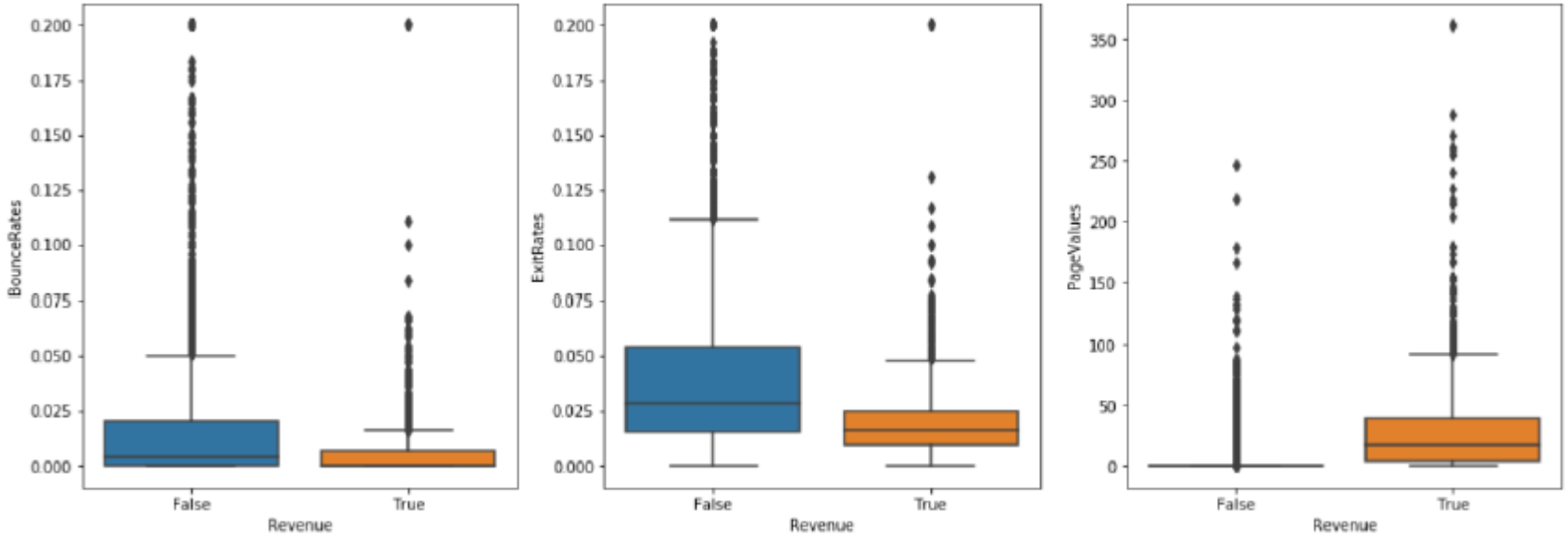


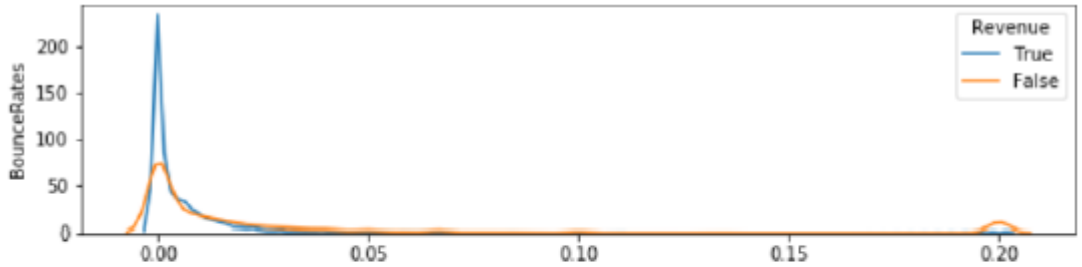
Most Used Browser Type

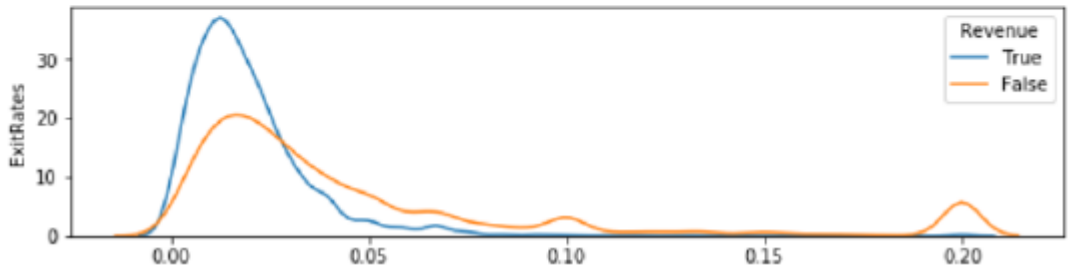
Highest Revenue Generation

Region	Type 1
Traffic	Type 2
Operating System	Type 2
Browser	Type 2

Insights: Impact of Bounce rates, exit rates and Page Values



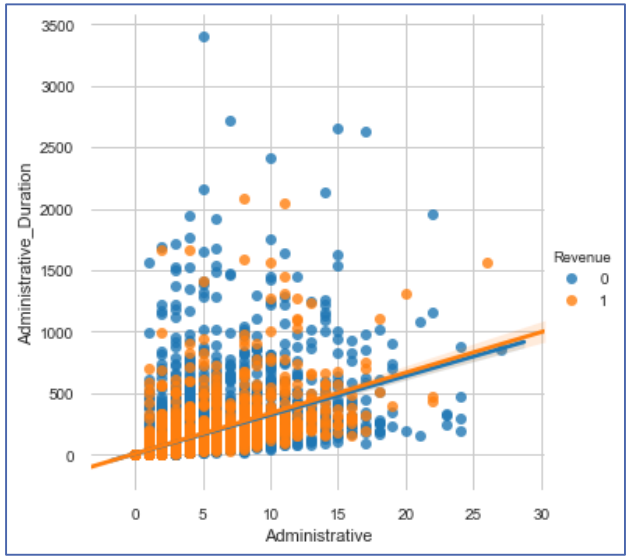


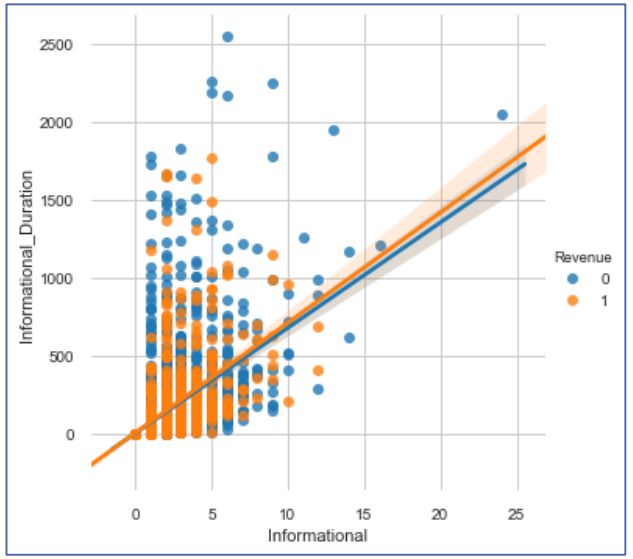


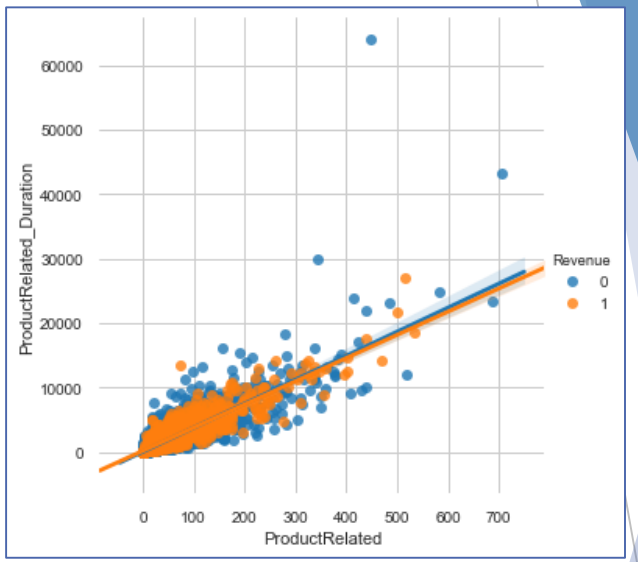
Insights

- For effective revenue generation low bounce rates, low exit rates and high page values hold the key.
- Focus on this combo can give good dividends as part of business strategy.

Insight & Correlation Study







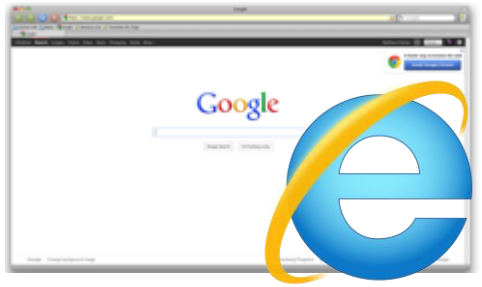
Insights

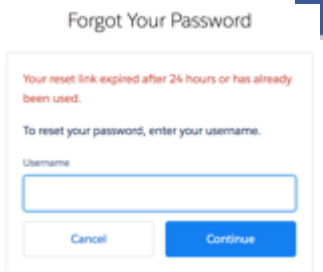
- Administrative and Administrative Duration, Informational and Informational Duration, Product Related and Product Related Duration are positively correlated.
- On Administrative pages 2 to 15 (probably pages like login, logout, password recovery, profile, email wish list etc.), visitors have spent more than 500 seconds (approx. 8 minutes) which is generally quite higher than normal. It suggests that visitors are having trouble logging in or it's taking too much time to process the request.
- Even though customers/visitors have spent a large amount of time on Product related pages, but the revenue generation is very low. There are certain outliers who spent more than 30000 seconds (approx. 8 Hours) but still didn't make any transaction (possibly screen left idle).

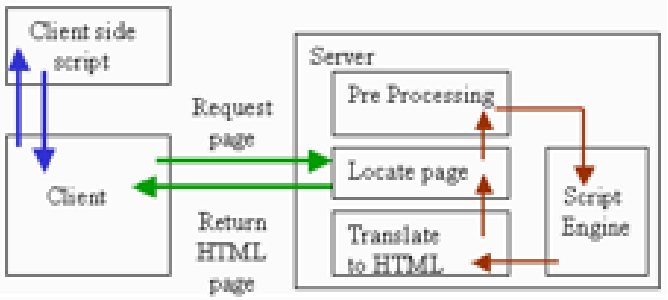
Recommendations - 1

- Number of Administrative and Information web pages on the website should be as low as possible.
- Product related web pages to be expanded as users are willing to spend more time on them.
- As part of business strategy focus on combination of low bounce rates, low exit rates and high page values
- Promotion Strategies should focus more on returning visitors, as they contribute more to revenue generation.
- Business should look to maximize the conversion of online visits to actual purchases in May and November.
- Target the audience from region 1 and 3 directly as these regions account for maximum revenue generation.

Recommendations – 1 (contd…)

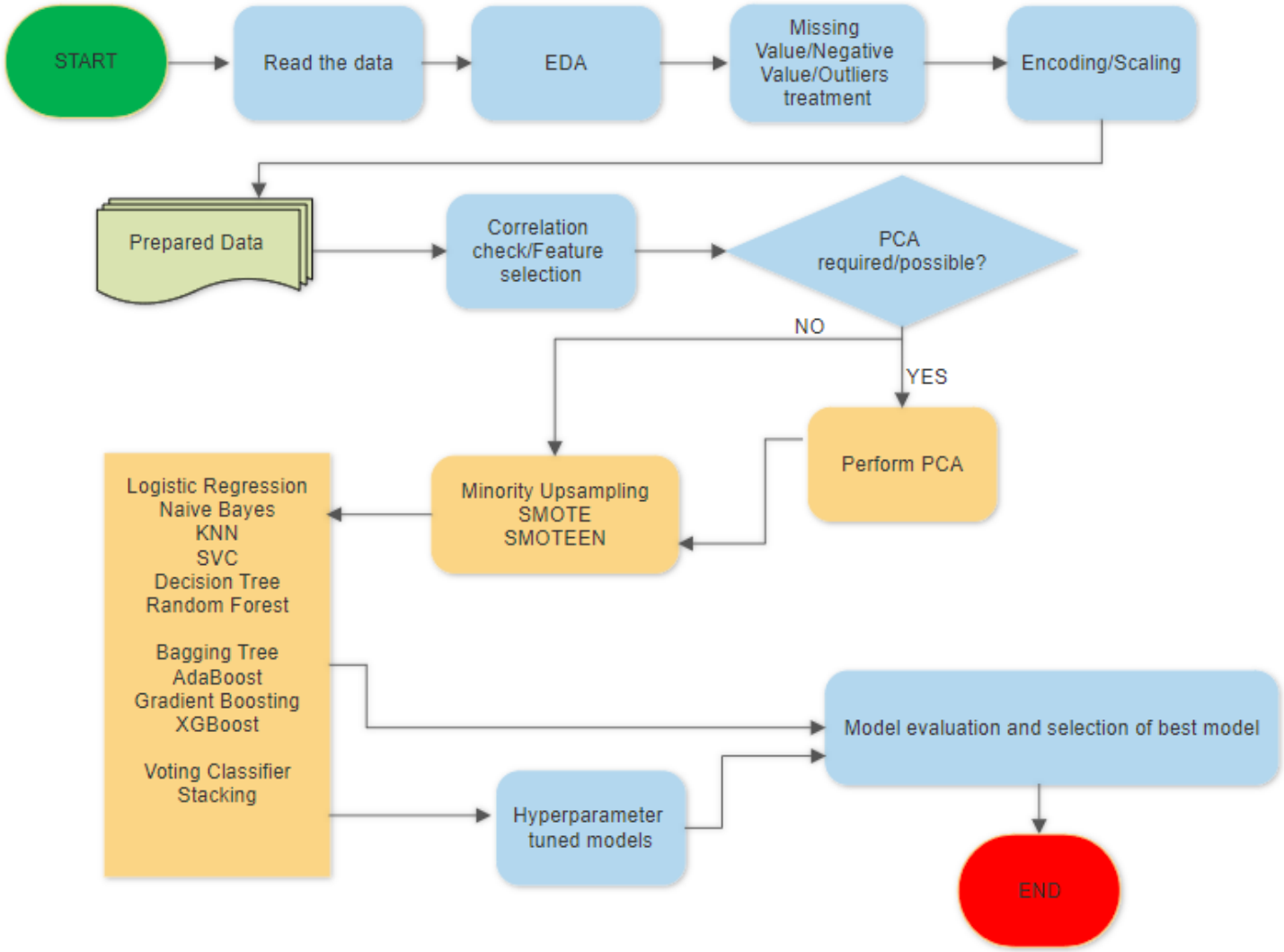






- Promote the OS type 2 and browser 1 and 2 on the website, something like – "For best results use OS x and browser y" etc.
- New visitors take up a larger percentage in those who complete purchase. So, it will be a good idea if business form marketing plans to attract new users every day.
- Administrative pages like login, logout, password recovery, profile, email wish list etc. need to be fixed.
- Automated client-side scripts should be embedded into the web pages so that dormant/idle sessions get logged out after some time of inactivity.

Modelling Approach



Handling class imbalance

Random Oversampling

- ✓ Replicate randomly the minority class examples.
- ✓ Ratio of two classes 1:1

SMOTE

- ✓ Select examples from the minority class to synthesize new examples with slightly perturbing feature values.
- ✓ Ratio of two classes 63:38

SMOTEEN

- ✓ Oversampling of minority class using SMOTE followed by under-sampling the majority class using ENN.
- ✓ Ratio of two classes 7:3.

Modelling Results

MODEL	Train Accuracy	Test Accuracy	True Positive (Non-revenue)	True Negative (Revenue)	CV Accuracy	CV Precision	CV Recall	CV F1 Score	CV ROC AUC	Remarks
Logistic Regression	91	87	2817	417	91	93	76	84	87	
Naïve Bayes	87	77	2403	472	87	75	84	79	86	
KNN	100	80	2551	393	95	94	91	92	94	Overfit
SVM	100	83	2767	318	96	97	90	93	95	Overfit
AdaBoost	94	88	2833	422	93	92	84	87	90	
Gradient Boosting	100	87	2795	437	94	92	89	90	92	
Bagging Tree	91	88	2827	413	91	91	78	84	87	Close to best
Decision Tree	79	81	2747	254	79	70	57	62	73	
Random Forest	100	88	2796	437	95	92	90	91	93	
Voting Classifier	99	85	2704	437	95	93	90	92	94	
XGBoost	95	88	2809	428	94	92	87	89	92	Best

Strategy to select best model ?

- ↘ We removed overfitted models.
- ↘ Looked for models which can predict good number of True Negatives i.e. Revenue generating user sessions. Our aim was to find out the users who are giving Revenue even if that means reducing True Positives i.e. Non-revenue generating user sessions..
- ↘ Judge models based on cross validation metrics.
- ↘ F1 Score which is the harmonic mean of Precision and Recall will be our best indicator as it is a general rule to look for higher F1 Score if our aim is to predict the minority class which here is Revenue generating user sessions more so in imbalanced data.
- ↘ After F1 Score, we looked for cross validation accuracy and cross validation ROC AUC metrics.

Recommendations - 2

- Attributes to be focussed upon for accurate prediction - ***PageValues, Month, ProductRelated_Duration, ExitRates, TrafficType, Administrative_Duration, Administrative, Informational_Duration, VisitorType, OperatingSystems, Weekend, SpecialDay, Region, Browser***
- Focus on improving mobility between pages to encourage users to browse among different products.
- Capitalize in the months of May and November. Provide additional deals/discounts to encourage product sales.
- Invest on ensemble machine learning classifiers as they proved more efficient in this case and could predict well on unseen data.
- Invest on data collection to improve success rates, scalability of ML algorithms and reduce bias due to class imbalance.

THANK YOU

APPENDIX : Feature Selection techniques used

- Information Gain (Mutual information & SelectKBest)
- Fisher Score (Categorical Variables)
- Univariate ROC_AUC
- Step Forward, Step Backward and Exhaustive Feature Selection
- Random forest feature importance
- Random forest recursive feature elimination
- Feature shuffling
- Hybrid recursive feature elimination(XGBoost)
- Hybrid recursive feature addition(XGBoost)
- Gradient boosting importance

APPENDIX : Other modelling results

MODEL	Train Accuracy	Test Accuracy	True Positive (Non-revenue)	True Negative (Revenue)	CV Accuracy	CV Precision	CV Recall	CV F1 Score	Remarks
Logistic Regression	89	88	3059	204	88	74	38	50	
Random Forest	99	90	3038	281	89	74	48	58	
Decision Tree	100	86	2880	315	86	54	56	55	Overfit
Naïve Bayes	36	36	802	547	35	19	96	31	Worst
KNN	90	87	3005	211	88	68	38	49	
SVM	88	88	3072	179	88	77	33	46	
AdaBoost	90	89	2972	316	89	68	58	62	
XGBoost	92	90	2979	347	91	75	62	68	Best
Gradient Boosting	92	89	2971	338	90	73	61	66	

Models on Base data

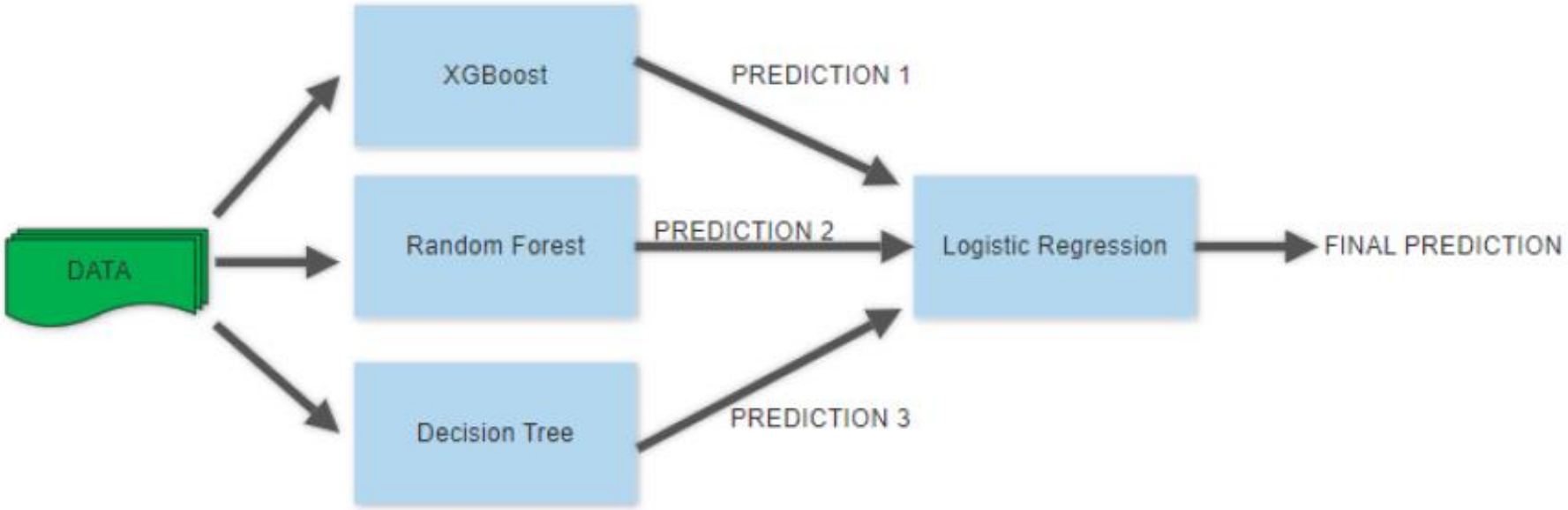
MODEL	Train Accuracy	Test Accuracy	True Positive (Non-revenue)	True Negative (Revenue)	CV Accuracy	CV Precision	CV Recall	CV F1 Score	CV ROC AUC	Remarks
Logistic Regression	72	72	2117	536	82	85	76	80	90	
Naïve Bayes	70	71	2120	503	75	74	77	76	83	
KNN	83	83	2522	560	89	83	99	90	90	
SVM	87	87	2736	469	95	92	98	95	95	
AdaBoost	36	38	859	559	86	87	84	85	86	
Gradient Boosting	39	40	972	501	93	90	97	93	93	
Bagging Tree	49	50	1340	519	95	92	99	95	95	
Decision Tree	32	33	666	543	70	68	80	73	70	Worst
Random Forest	52	54	1449	548	95	92	99	96	95	
Voting Classifier	81	80	2426	549	96	93	99	96	96	Best
XGBoost	39	41	998	501	93	90	97	93	93	

Models on upsampled data

MODEL	Train Accuracy	Test Accuracy	True Positive (Non-revenue)	True Negative (Revenue)	CV Accuracy	CV Precision	CV Recall	CV F1 Score	CV ROC AUC	Remarks
Logistic Regression	83	88	2904	368	83	85	66	74	80	
Naïve Bayes	75	73	2248	445	75	63	79	70	76	
KNN	100	81	2647	358	90	80	96	87	91	Overfit
SVM	99	85	2959	186	94	91	94	92	94	Overfit
AdaBoost	92	88	2916	351	90	88	85	86	89	
Gradient Boosting	97	89	2937	366	92	90	89	89	91	
Bagging Tree	87	87	2799	426	86	84	80	81	85	
Decision Tree	78	81	2746	268	76	72	58	64	72	
Random Forest	100	89	2910	376	93	89	91	90	92	
Voting Classifier	97	88	2948	321	91	90	84	87	89	
XGBoost	91	88	2874	402	91	88	87	87	90	Best

Models on SMOTE data

APPENDIX : Stacking of Models attempt

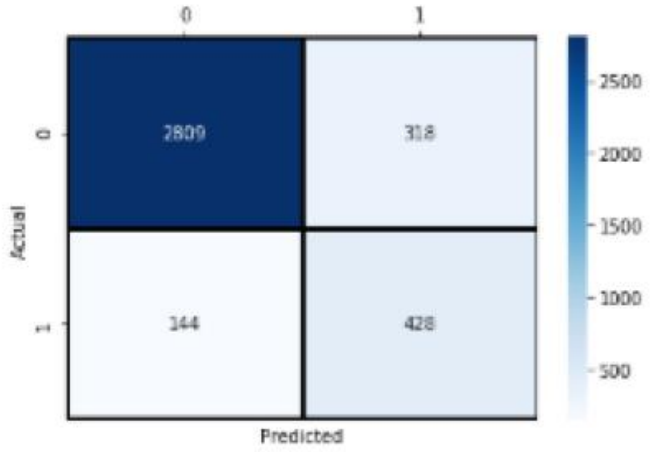


APPENDIX : Best “XGBOOST” Model Performance Metrics

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=200, n_jobs=-1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
Training Accuracy : 0.9507169538498614
Testing Accuracy : 0.8751013787510138
ROC AUC Score : 0.8232784164987555
```

****

CONFUSION MATRIX



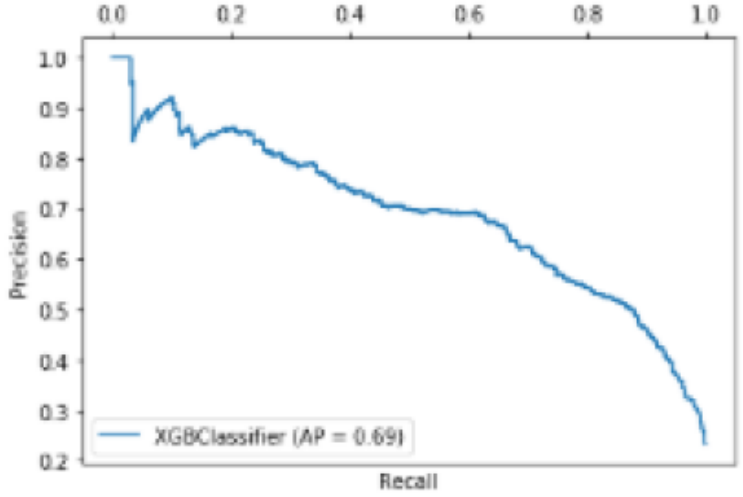
****

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.95	0.90	0.92	3127
1	0.57	0.75	0.65	572
accuracy			0.88	3699
macro avg	0.76	0.82	0.79	3699
weighted avg	0.89	0.88	0.88	3699

****

PRECISION RECALL CURVE



CROSS VALIDATION METRICS

```
Mean Accuracy : 0.9361392009533913
Standard Deviation : 0.023423236706950895
Mean precision score : 0.9223367967239311
Standard Deviation precision score : 0.020356141346406367
Mean recall score : 0.8651577818627452
Standard Deviation recall score : 0.07517636773543315
Mean f1 score : 0.8912494391510613
Standard Deviation f1 score : 0.04348419086508033
Mean AUC ROC Score : 0.9163931614964383
Standard Deviation AUC ROC Score : 0.037504555687855004
```

****