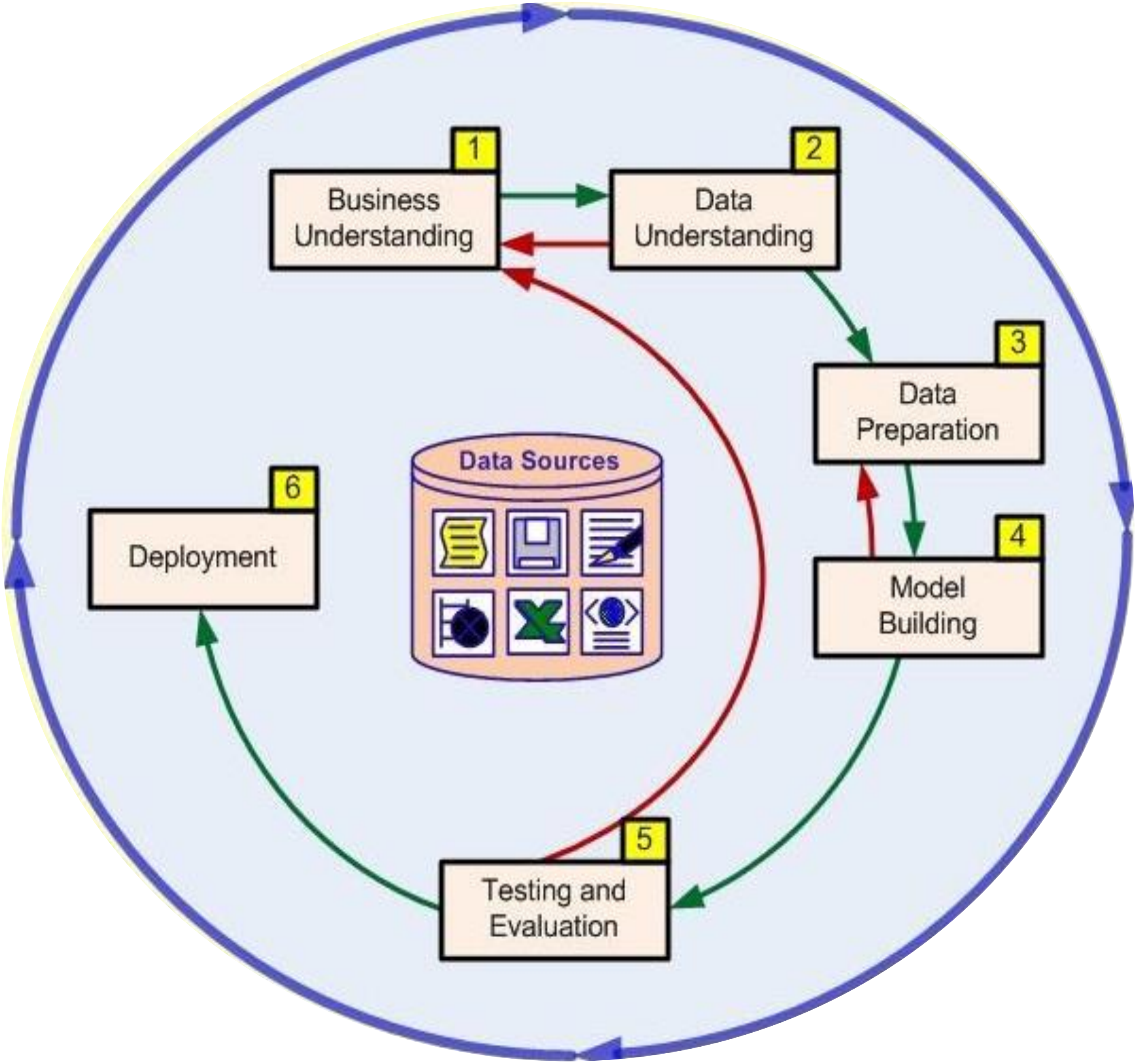
Online Shopper Purchasing Intention

Presented by Atlanta Rapid Miners

Our Process: The **CRISP** Data Mining Framework



Business problem: How can we engage our customers based on their purchase intention to maximize profit?

The problem in e-commerce

Low Conversion Rate

- E-commerce websites today tend to have a very low conversion rate of 5-15%
- Traffic is not usually a problem: More visitors search for different products without ever buying them

Customer Retention

- With increasing online competition, customers tend to compare products across platforms and purchase those with a mix of low prices and high ratings

Our solution

Predict customer intention

- Predict if a customer is going to buy or not as soon as they visit the website and target them while they are still in the session as per their segment

Why target while still in session?

- Companies tend to e-mail non-purchasing customers a few hours (or days) after they end their session with offers
- A delay in messaging can cause customers to lose interest. Providing offers (almost) instantly, we can convert customers who did not originally intend to purchase

Business Understanding → Data Understanding → Data Preparation → Modelling → Evaluation → Deployment

Segmenting our customers based on their Intention

Intention Buy

Barbara

Can be casual or frequent shopper

- Routinely checks favorite shopping sites with a plan to purchase
- May or may not have a specific product in mind

Marketing Strategy

Upselling

Cross-Selling

AOV Increase

Goal: We want to encourage this segment to purchase more!

Intention Browse

Patrick

Explores but doesn't purchase

- Frequently browsing on the website but never makes a purchase
- Tends to have a high bounce rate or exit rate

Marketing Strategy

Cross-Channel Strategy

Impulse Purchase

Goal: We want to encourage this segment to make a purchase after browsing

Data Understanding

Data Source and Description

- Online Shoppers' Purchasing Intention dataset provided by UCI Machine Learning Repository
- Includes 12,330 instances of customer characteristics, behaviors, and purchase decisions. Each instance belongs to a different customer.
- Contains 18 attributes: 10 numerical and 8 categorical.
- There are no missing values or anomalies (such as negative values)
- The **Revenue** attribute is used as the target variable: FALSE refers to **No Purchase**, TRUE refers to **Purchase**

Independent Variables

Page Type and Duration (Continuous)

Informational pageviews
Informational duration
Product pageviews
Product page duration

Google Analytics Metrics (Continuous)

Bounce Rate
Exit Rate
Page Value (in USD)

Time Information

Special Day (Continuous)
Month (Categorical)
Weekend (Categorical)

Visitor Characteristics (Categorical)

Operating Systems
Browser
Region
Traffic Type
Visitor Type

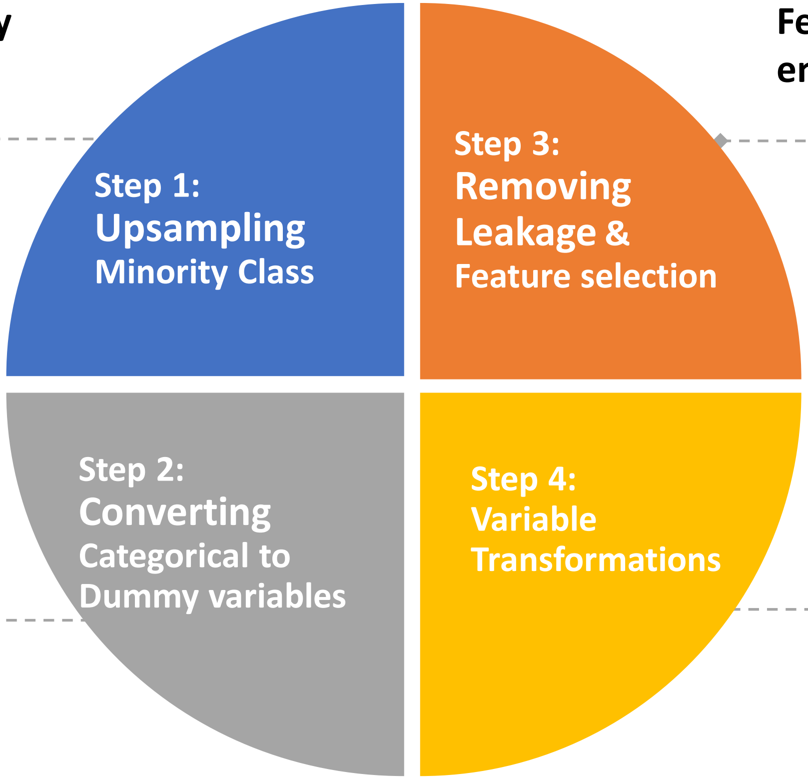
Ideal but Unavailable Information

- Information on multiple sessions for each customer, so that it is possible to analyze the willingness to purchase at different timeframes of each customer's journey

Data Preparation and Cleaning

Up-sampling minority class

Our target variable, Revenue, is slightly imbalanced. We have used **randomized** up-sampling to change the distribution **by up-sampling the minority class** (From 85%-15% to 70%-30%)



Feature selection and engineering

We have **excluded leakage variables** such as **duration** and **pageviews** and will restrict our predictions to the time point when user starts the journey.
We also took the **average of Bounce rate and Exit rate** as they represented similar information in most cases

Dummy variables for non-binary categories

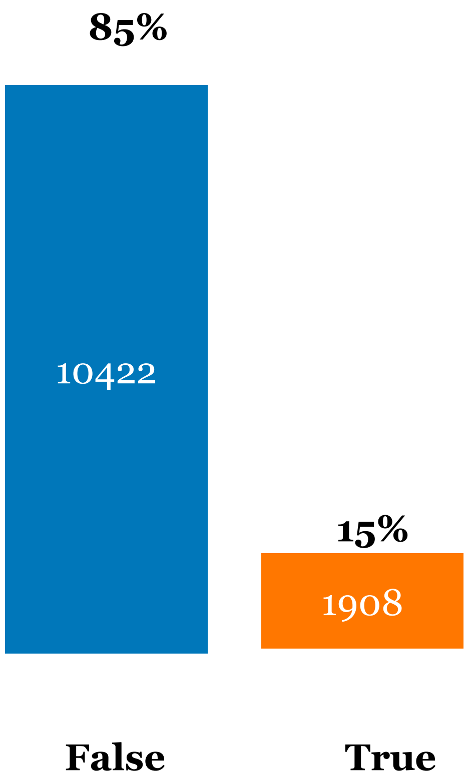
Our categorical variables are unordered. It is imperative that we change them to multiple **dummy variables** using pandas.

Monotonic Transformations

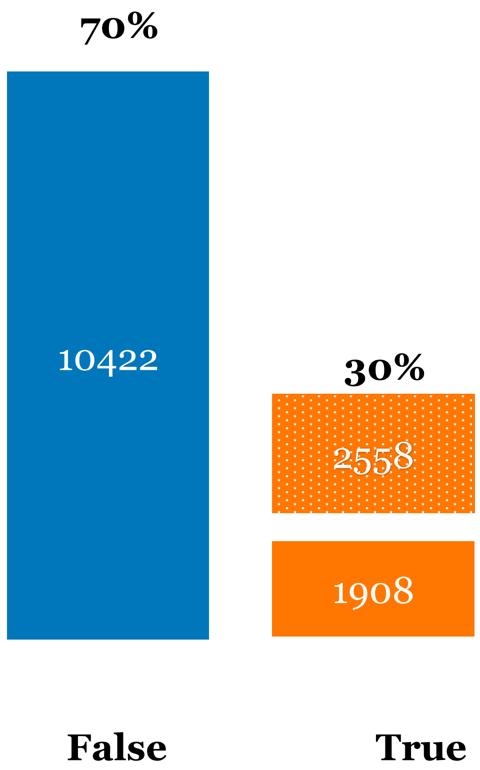
We tried our models **with and without** transformed variables. Variables that had a skewed distribution were transformed using logs,sq roots and cuberoots.
This was an iterative process which involved calibrating multiple models.

Data Preparation

Up-sampling the Minority Class

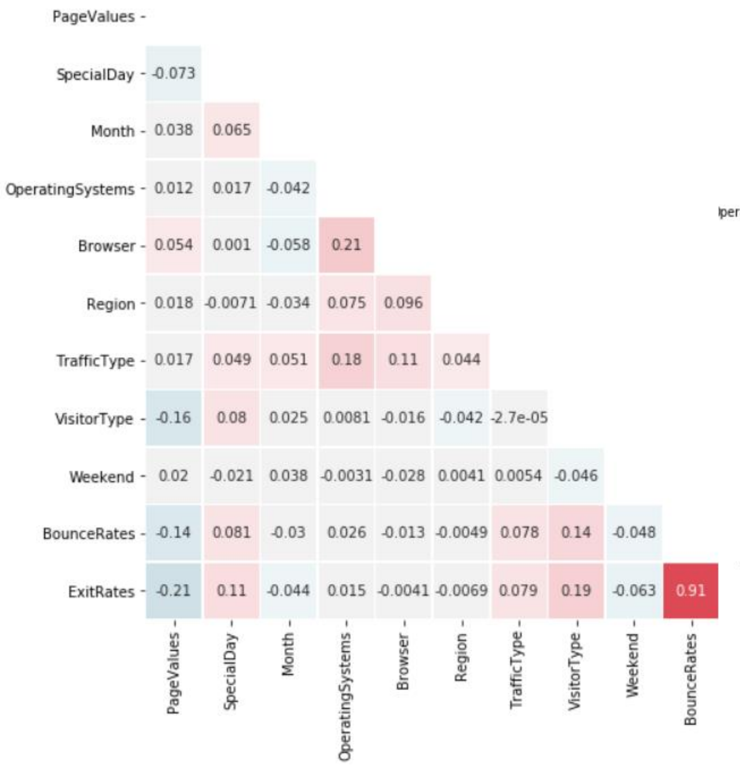


Before Up-sampling



After Up-sampling

Feature Engineering

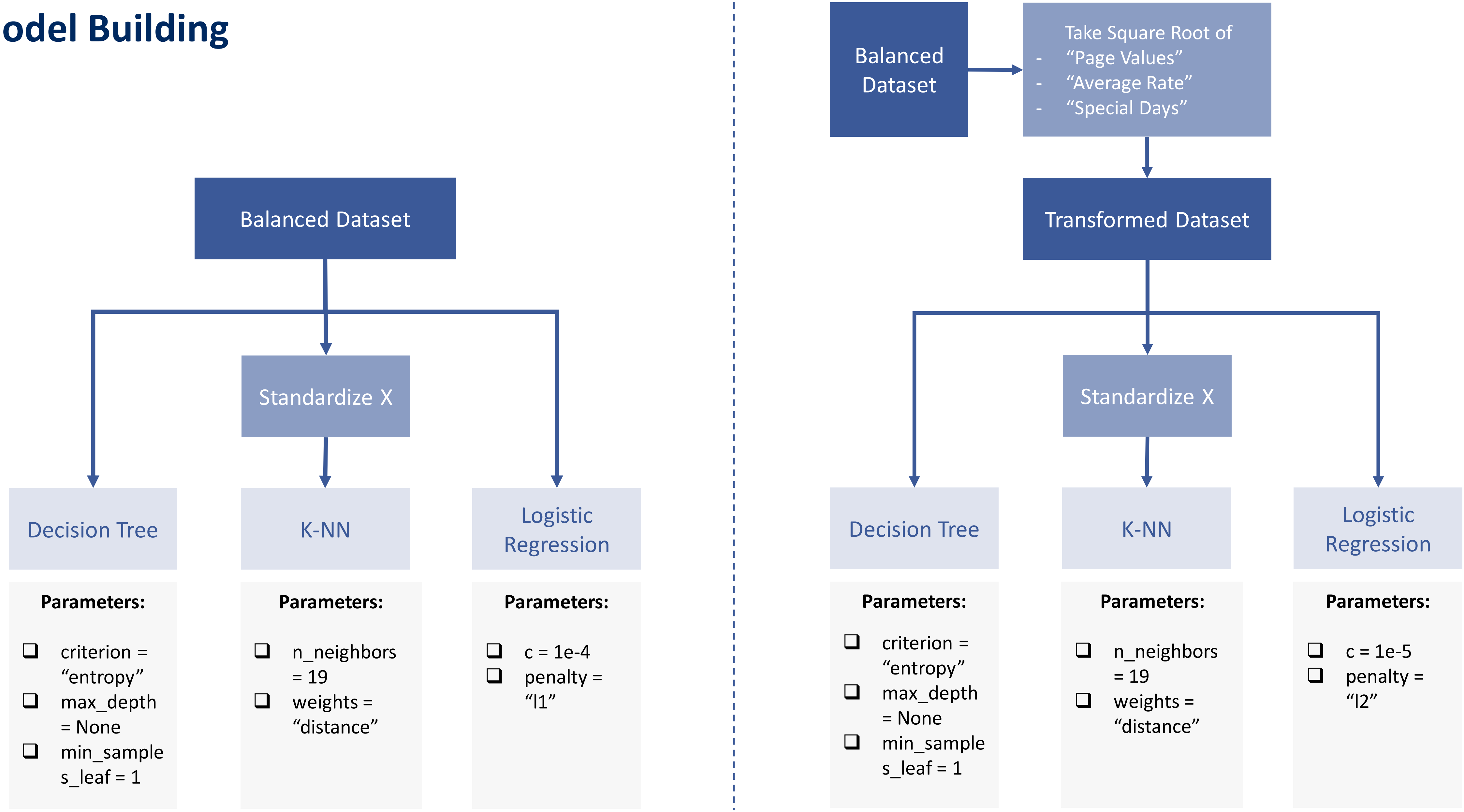


With Bounce and Exit rates

With Average Rate

Business Understanding › Data Understanding › Data Preparation › Modelling › Evaluation › Deployment

Model Building



Business Understanding > **Data Understanding** > **Data Preparation** > **Modelling** > **Evaluation** > **Deployment**

Model Evaluation using transformed variables – *At a glance*

Criterion	Decision Tree	Logistic Regression	kNN	Naive
Accuracy	0.90	0.86	0.91	0.7
F1-Score$_{avg}$	0.89	0.76	0.88	-
AUC	0.91	0.88	0.96	-
Computational Power	Medium	Low	High	None
Interpretability	High	Low	Medium	High

Business Understanding > Data Understanding > Data Preparation > Modelling > Evaluation > **Deployment**

Model Evaluation

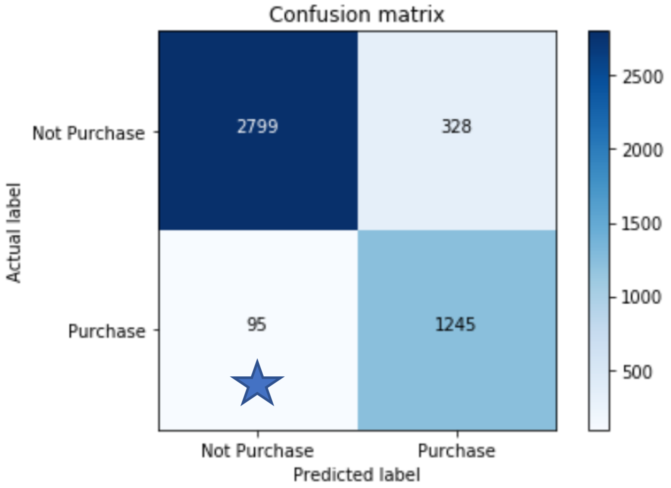
Decision Tree – Minimizes FNs

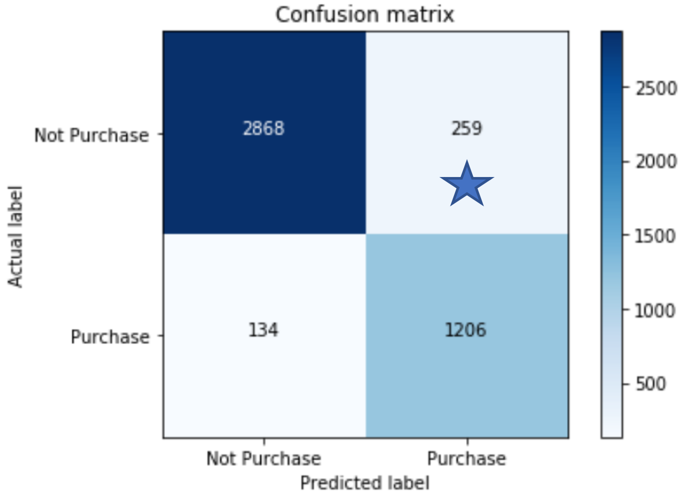
Accuracy: 90%
Precision: 88%
AUC: 91%
Recall: 91%
F-Measure: 89%

KNN – Minimizes FPs

Accuracy: 91%
Precision: 81%
AUC: 96%
Recall: 84%
F-Measure: 88%

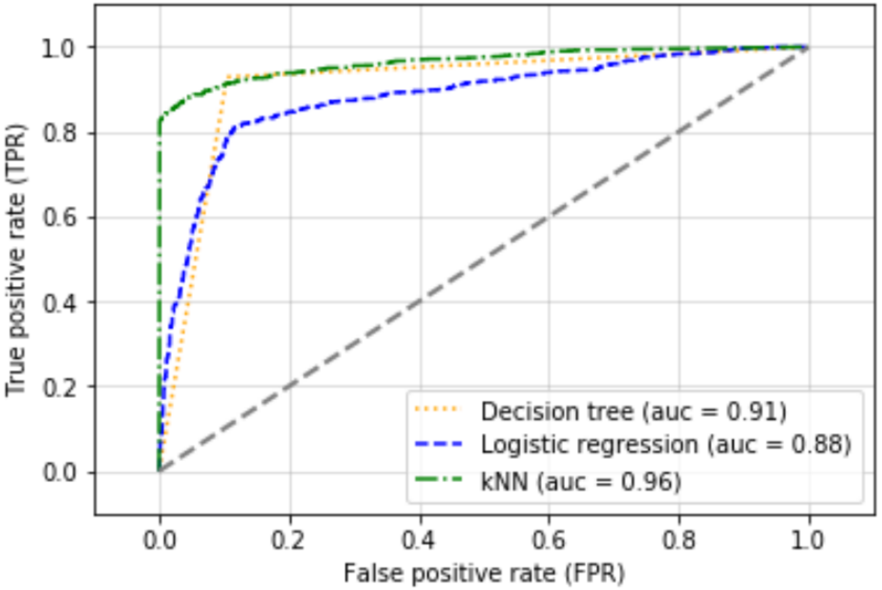
Confusion Matrix - Tree





Confusion Matrix - KNN

ROC Curve



So which model should we deploy?

Choosing the *ideal* model depends on the deployment plan, which in turn depends on which model can help us maximize revenue

Business Understanding › **Data Understanding** › **Data Preparation** › **Modelling** › **Evaluation** › **Deployment**

Sample Deployment Strategy and Expected ROI

Browsers – *Impulse Purchase Strategy*

- Offer discounts early in the session and keep a time cap: e.g. Purchase within the next 30 minutes!
- Engage customer & collect data: If un-availed, collect feedback on why a purchase wasn't made

Buyers – *Cross-selling and Upselling*

- Aim to increase Average Order Value (AOV) by
  - **Upselling**: Recommend more premium products
  - **Cross-selling**: Recommend additional products

ROI Analysis – *Show me the money*

The confusion matrix for DT on test data:

	No $Purchase_{Predicted}$	$Purchase_{Predicted}$
No $Purchase_{Actual}$	2799	328
$Purchase_{Actual}$	95	1245

Profit Matrix: Assumptions

- *Conversion rate for additional transactions :* ***10%***
- *Increase in AOV: $**50***
- *Cost involved in providing a discount that got availed: $**10***

	No $Purchase_{Predicted}$	$Purchase_{Predicted}$
No $Purchase_{Actual}$	**$40 (Impulse purchase)**	$0
$Purchase_{Actual}$	-$10	**$40 (Upselling/ Cross-selling)**

ROI using these strategies = [[(2799 * $40) + (1245 * $40)] * 10% conversion rate] - (95*10$) = $15,226

Business Understanding > Data Understanding > Data Preparation > Modelling > Evaluation > Deployment

Deployment Considerations and Risk Mitigation

Potential risks involved and how to eliminate/reduce them

Loss of Revenue

- **We lose money every time we predict a buyer as a browser.**
- By offering discounts to someone who would have purchased the product anyway, we lose money
- Our aim is to <u>preserve f1 score but also minimize FNs</u>
- **A PILOT program is necessary.**
  - Target only a percent of predicted *browsers*
  - Of those that are not targeted, collect data on outcome
  - Calibrate model

Loss of Customers

- Offering discounts and special offers is helpful in gaining a bigger customer universe
- In the long term, however, customers may get dependent on receiving special offers to make purchases
- Such strategies should be used in shorter terms to gain popularity, but should be combined with other strategies in the long term

Ethical Risks

- When looking at upselling or cross-selling for buyers, recommendations must be provided using unbiased ways
- For instance, recommending products that are authentic is important.
- Maximizing profit by cross-selling poor-quality products with high profit margins is neither ethical nor successful in long term

Process Recap

Business Understanding

- ✓ Why is conversion rate low?
- ✓ Who are our target customers?
- ✓ What can we do to increase their purchase intention?

Data Understanding and Preparation

- ✓ Data Source
- ✓ Data Types
- ✓ Up-sampling the Minority Class
- ✓ Data Selection
  - ✓ Remove Data Leakage
  - ✓ Feature Selection
  - ✓ Data Transformation

Modeling and Evalution

- ✓ Compare multiple classification models
- ✓ Minimize False Negatives (FNs) and preserve F-1
- ✓ Evaluate ROI if deployed

Business Deployment

- ✓ Solution to overcome low conversion rate?
- ✓ How to mitigate risk?

Thank You

Questions?

Appendix 1: Dataset: Before and after cleaning and transforming variables

Dataset after removing leakage variables

BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.20	0.20	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.00	0.10	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.20	0.20	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
0.05	0.14	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
0.02	0.05	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

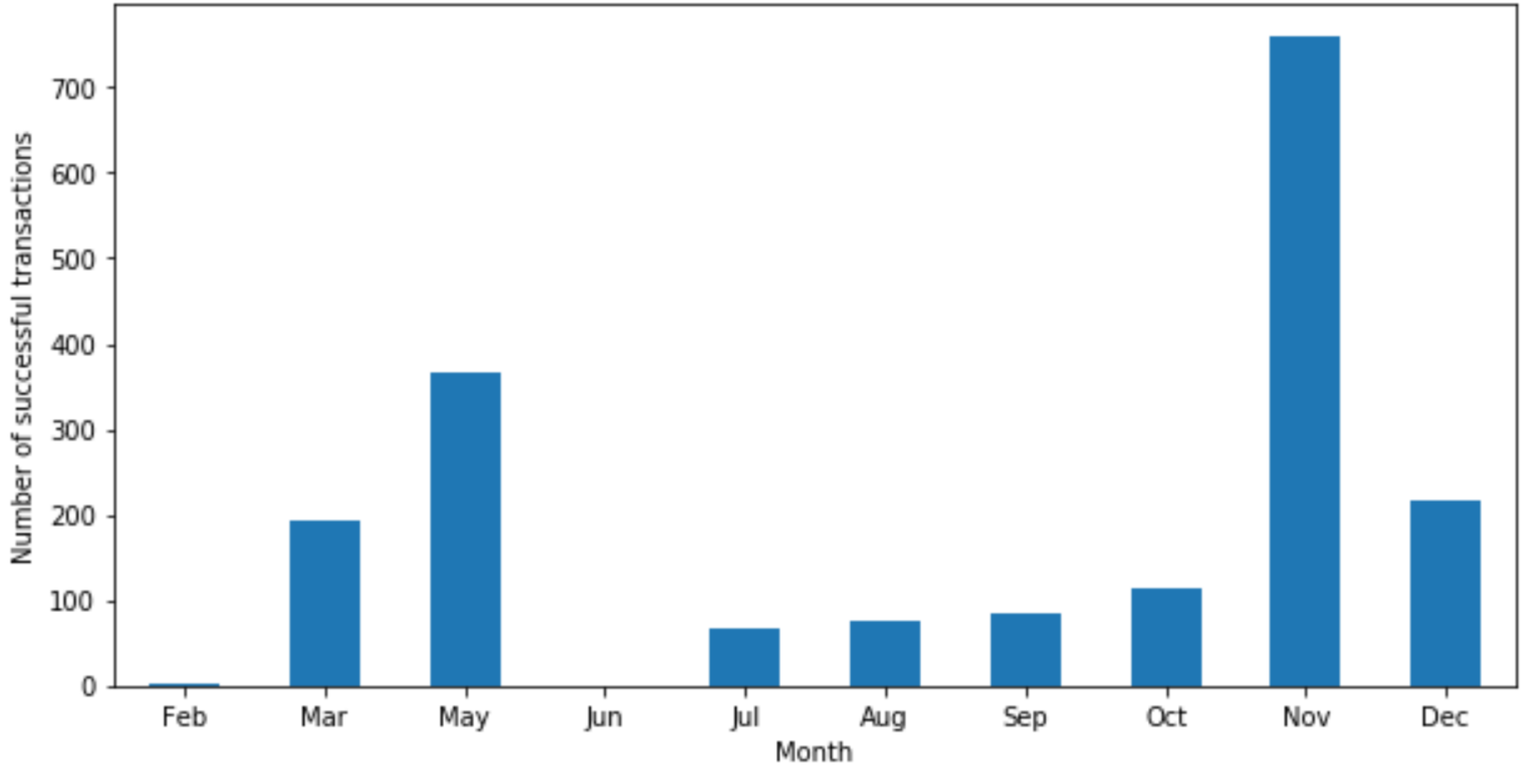
Dataset after using label encoder to transform text categories to numbers

BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.20	0.20	0.0	0.0	2	1	1	1	1	2	0	0
0.00	0.10	0.0	0.0	2	2	2	1	2	2	0	0
0.20	0.20	0.0	0.0	2	4	1	9	3	2	0	0
0.05	0.14	0.0	0.0	2	3	2	2	4	2	0	0
0.02	0.05	0.0	0.0	2	3	3	1	4	2	1	0

Dataset after transforming 3 variables to square root and taking average of Bounce Rate and Exit rate as Average rate

Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue	PageValues_sqrt	Average Rate_sqrt	SpecialDay_sqrt
2	1	1	1	1	2	0	0	0.0	0.447214	0.0
2	2	2	1	2	2	0	0	0.0	0.223607	0.0
2	4	1	9	3	2	0	0	0.0	0.447214	0.0
2	3	2	2	4	2	0	0	0.0	0.308221	0.0
2	3	3	1	4	2	1	0	0.0	0.187083	0.0

Appendix 2: Sales by month

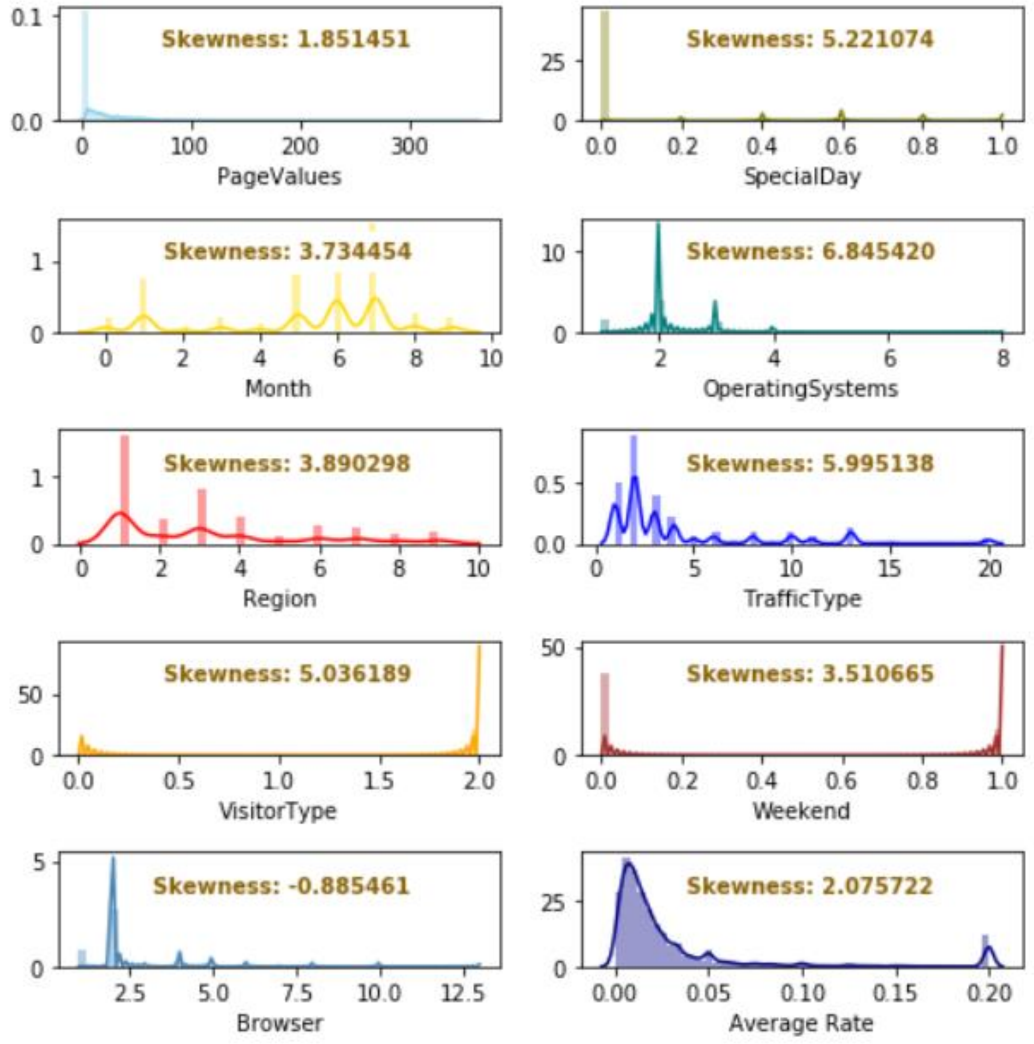


Insights:

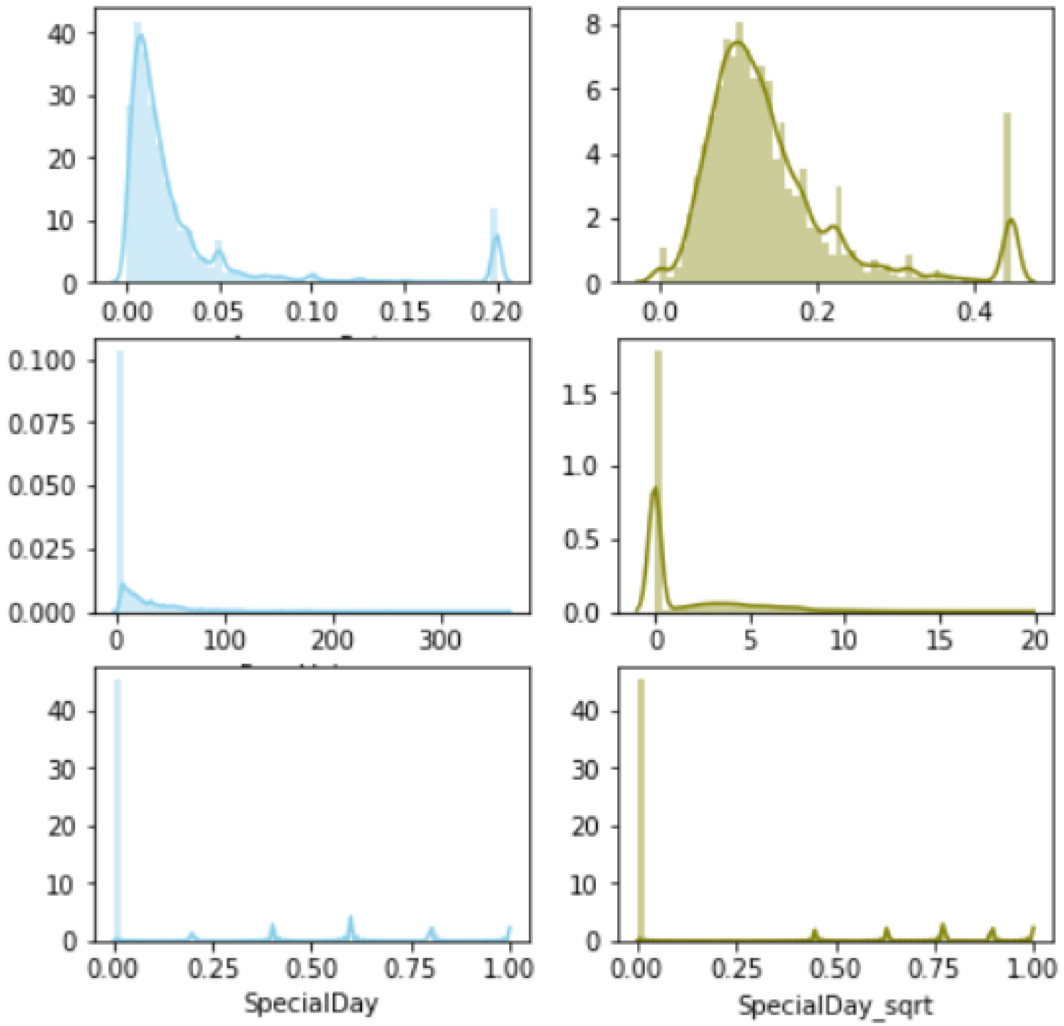
- Highest number of sales in the month of November, while low sales in Feb and June
- It is important to note that we have data for 10 months. i.e. we are missing data for Jan and April

Appendix 3: Skewness and Data Transformation

Skewness of the Prepared Dataset

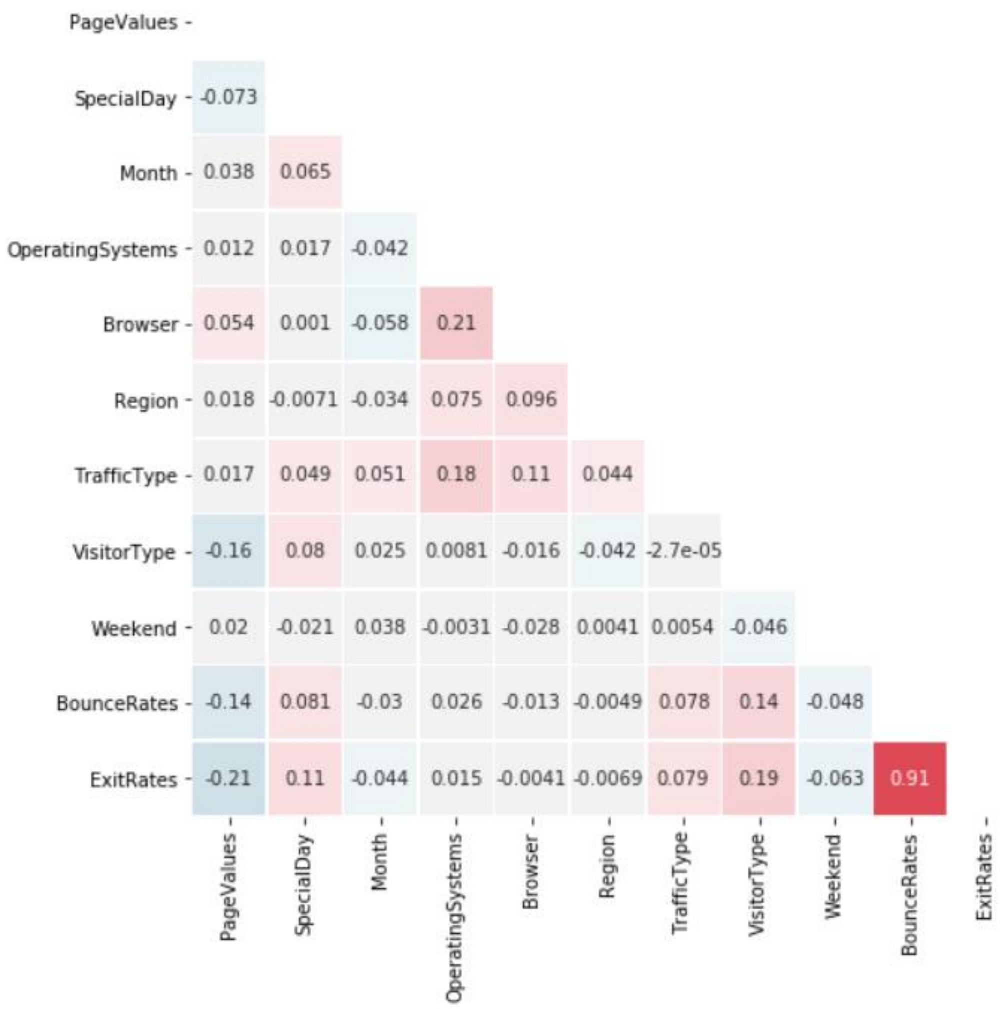


Skewness of the Transformed Dataset

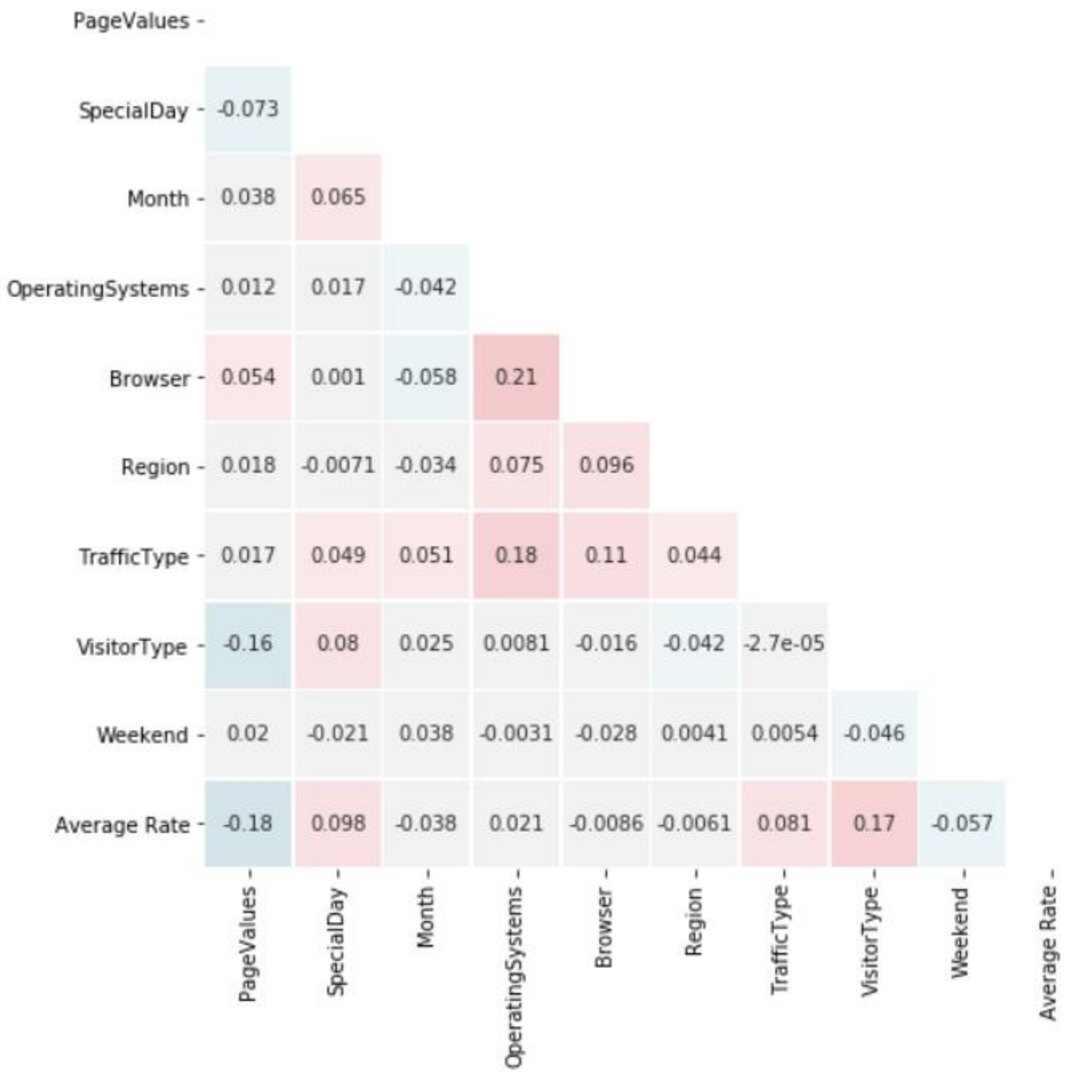


Appendix 4: Correlation Heatmap

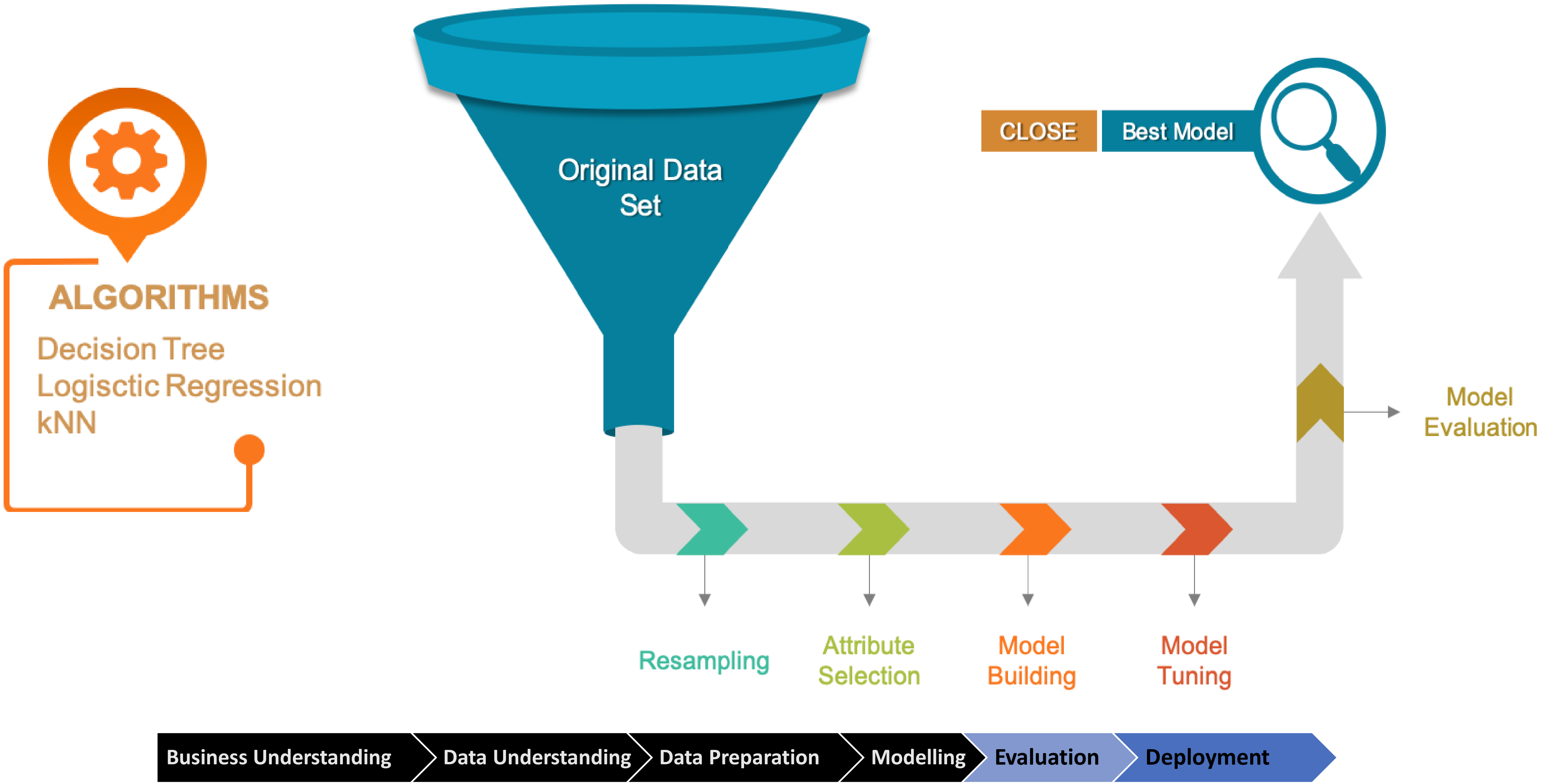
Correlation table the Original Dataset



Correlation table after feature engineering



Appendix 5: Modeling Process

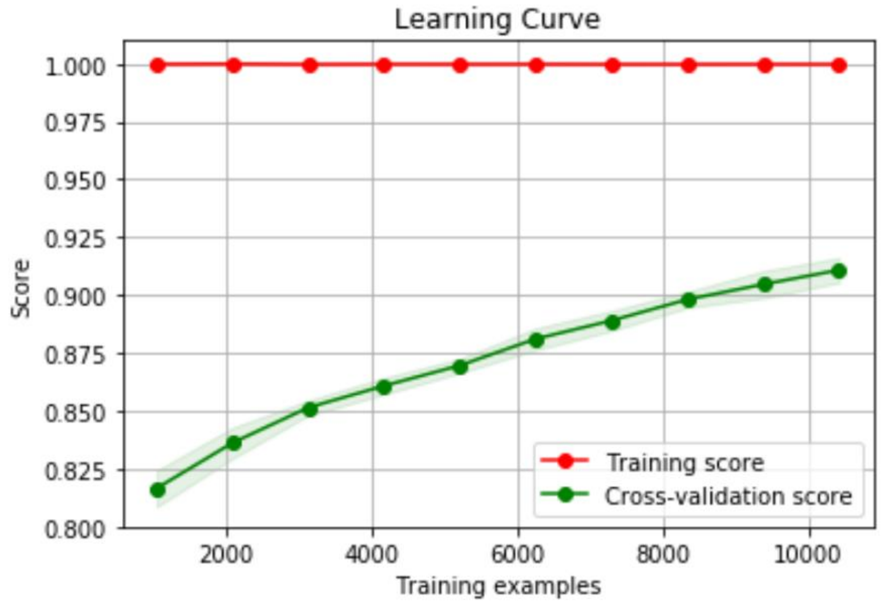


Appendix 6: Modeling Evaluation – Learning Curve and Fitting Graph

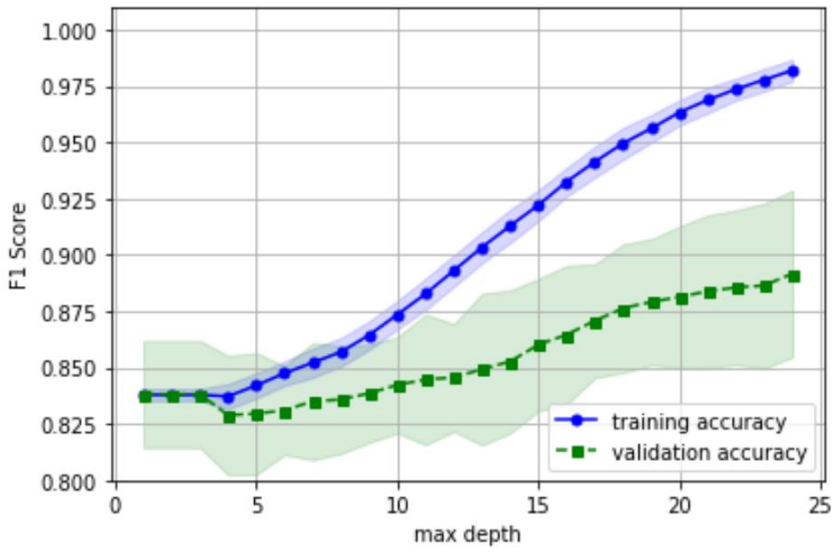
Generalization Performance

The accuracy rate for training set remain 1 because the max depth is not limited. We could observe the gap is shrinking as the training example increases, suggesting there might be more improment from additional data.

After reaching tree_depth 25, the model produces F1 score close to the optimized model, and we may consider stop adding depth in case the overfitting problem is worsen.



Learning Curve



Fitting Graph

Business Understanding → Data Understanding → Data Preparation → Modelling → Evaluation → Deployment