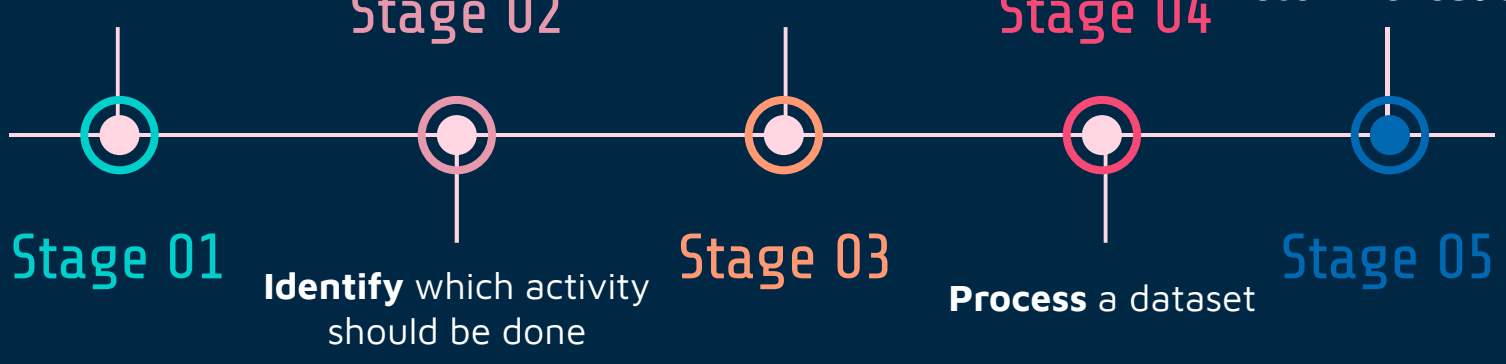
DATA SCIENCE 8
PROJECT PRESENTATION

OUR PROCESS


Stage 01
Understand dataset
Stage 02
Identify which activity should be done
Stage 03
Analyze data with exploration and visualization
Stage 04
Process a dataset
Stage 05
Develop model which relevant with the problem along with **evaluation** and **recommendation**

DigitalSkola
PepeThink

Understand

01

HackerEarth : How not to lose a customer in 10 days

This Dataset belongs to a Machine Learning Challenge hosted at Hacker Earth.

To **estimate** the churn risk rate of each customer.

Churn risk rate is a marketing metric that describes the **number of customers who leave a business** over a specific time period.

Every data is assigned a prediction value that estimates their **state of churn** at any given time, based on:

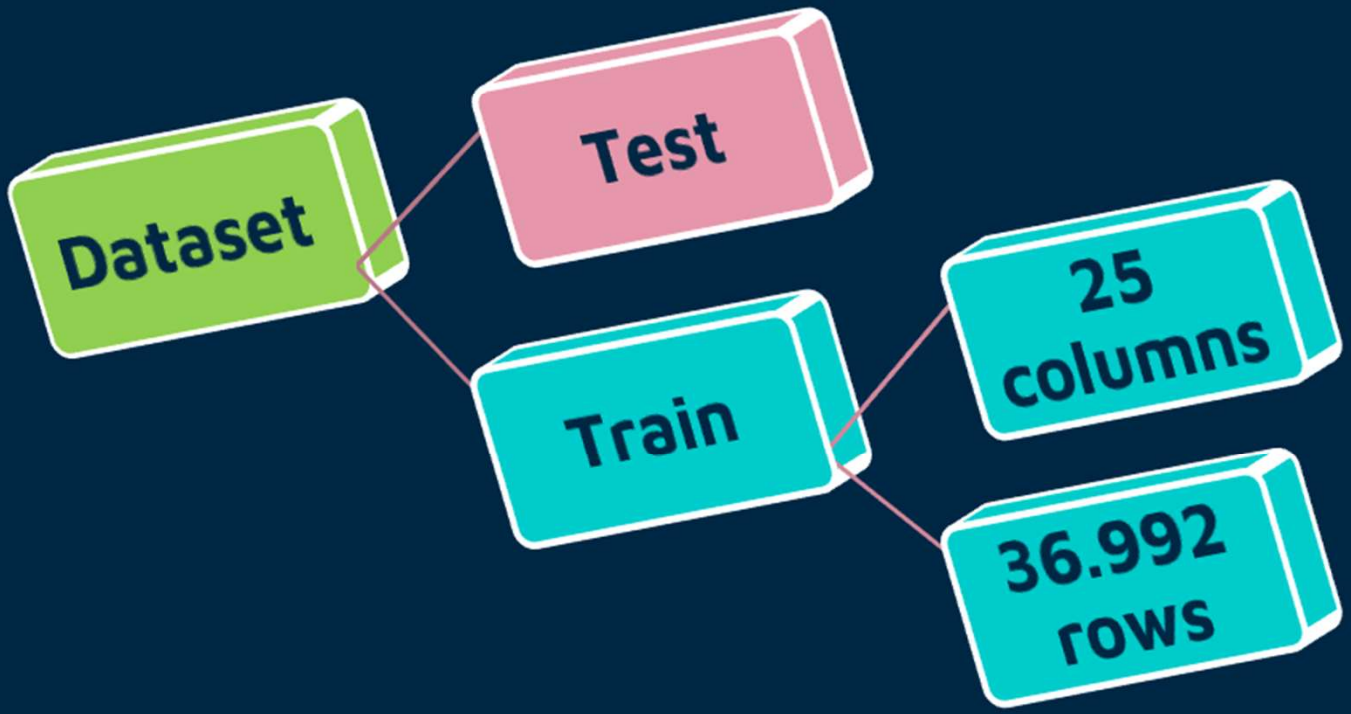
Customer Information

Browsing Behavior

Historical Purchase

DigitalSkola
PepeThink

HackerEarth : The Dataset



24

Features

1

Target

Customer Information 10	Browsing Behavior 6	Historical Purchase 8	Churn
Customer_id	Medium_of_operation	Preferred_offer_types	Churn_risk_score
Name	Internet_option	Avg_transaction_value	
Age	Last_visit_time	Points_in_wallet	
Gender	Day_since_last_login	Used_special_discount	
Security_no	Avg_time_spent	Offer_application_preference	
Region_category	Avg_frequency_login_days	Past_complaint	
Membership_category		Complaint_status	
Joining_date		Feedback	
Joined_through_referral			
Referal_id			

Target

Churn
Churn_risk_score

Our target is **churn_risk_score** column on the train dataset.

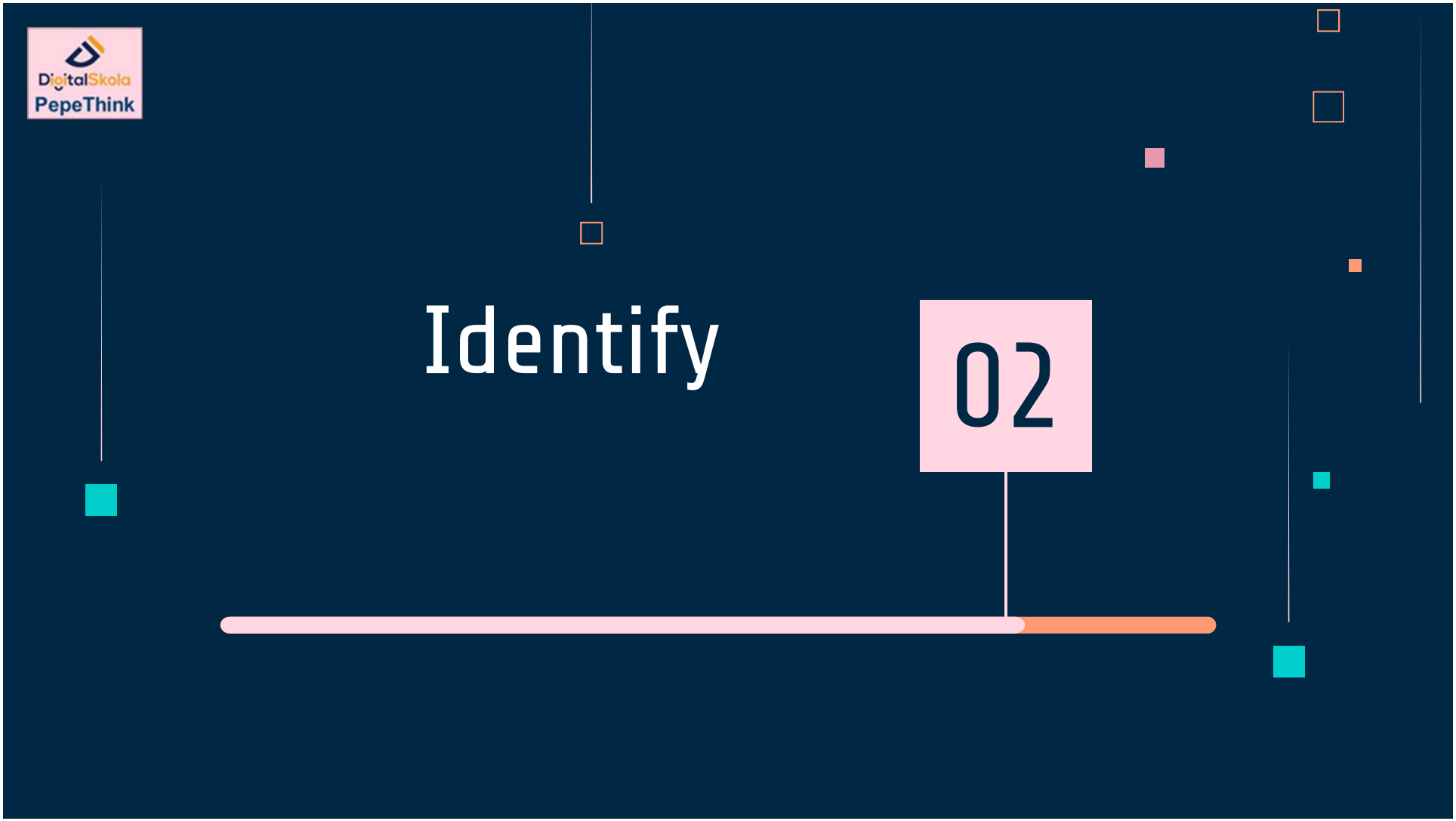
We will assess customers who have a churn risk level with a value between 1-5.

Value 1 has the smallest customer's probability to churn.

Value 5 has the biggest customer's probability to churn.

Note: Value of -1 in the churn column will not be used in the creation of the model so it will ignored.

DigitalSkola
PepeThink


DigitalSkola
PepeThink
Identify
02

Feature Data

We divide the data into two groups: numerical and categorical

Numerical

- age
- joining_date
- last_visit_time
- days_since_last_login
- avg_time_spent
- avg_transaction_value
- avg_frequency_login_days
- points_in_wallet

Categorical

- churn_risk_score
- customer_id
- Name
- gender
- security_no
- region_category
- membership_category
- joined_through_referral
- referral_id
- preferred_offer_types
- medium_of_operation
- internet_option
- used_special_discount
- offer_application_preference
- past_complaint
- complaint_status
- feedback

Error Data

Null Value

region_category	preferred_offer_types	points_in_wallet
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN

- region_category has 5263 (14.7%) null values
- preferred_offer_types has 276 (0.8%) null values
- points_in_wallet has 3341 (9.3%) null values

Error Data

Unknown and "xxxxxx" Value

referral_id	gender
xxxxxxxx	Unknown
xxxxxxxx	Unknown
xxxxxxxx	Unknown
xxxxxxxx	Unknown
xxxxxxxx	Unknown

- referral_id has 17296 (48.3%) "xxxxxx" values
- gender has 59 (0.2%) unknown values

Error Data

Minus Value

days_since_last_login	avg_time_spent	avg_frequency_login_days	points_in_wallet
-999	-1837.00837	-15.62939639	-760.6612363
-999	-643.0468413	-1.038410928	-549.3574977
-999	-882.8947139	-7.352443307	-506.2567158
-999	-371.6941047	-24.70043508	-469.0203988
-999	-20.34910768	-19.61911847	-445.2884572

- days_since_last_login has 1944 (5.4%)"-" values
- avg_time_spent has 1659 (4.6%)"-" values
- avg_frequency_login_days has 659 (1.8%) "-" values
- points_in_wallet has 134 (0.4%) "-" values

Error Data

"?" Value

joined_through_referral	medium_of_operation
?	?
?	?
?	?
?	?
?	?

- joined_through_referral has 5292 (14.8%) "?" values
- medium_of_operation has 5230 (14.6%) "?" values

Error Data

"Error" Value

avg_frequency_login_days
Error
Error
Error
Error
Error

- avg_frequency_login_days has 3419 (9.5%) "Error" Values

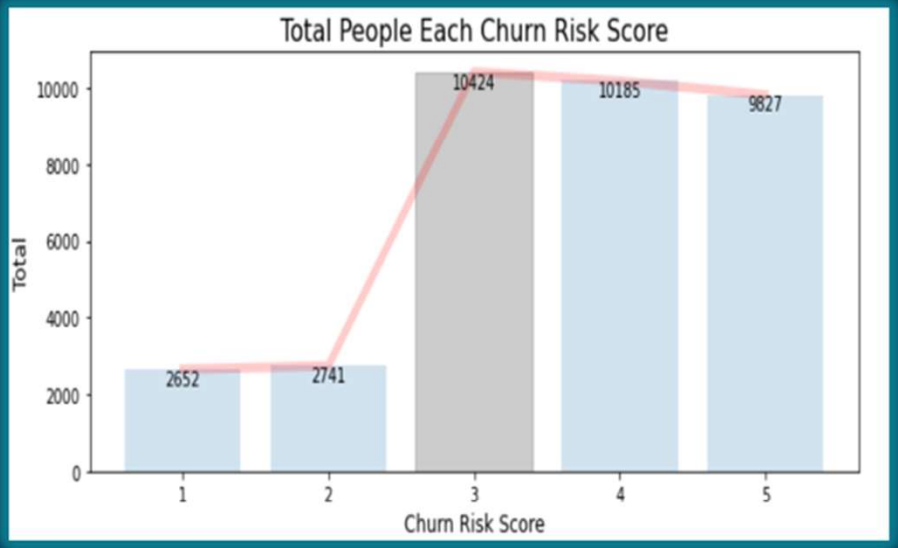
DigitalSkola
PepeThink

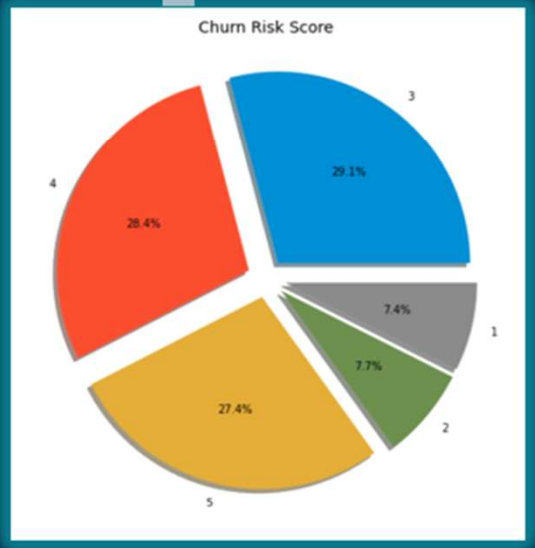
Analyze

Exploration and Visualization Data

03

Churn Risk Score



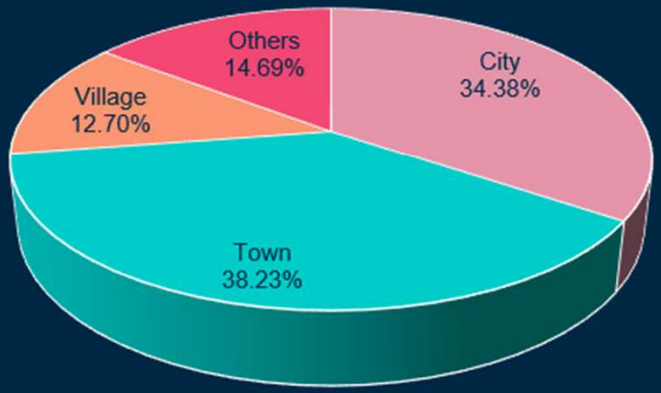


Churn_risk_score :

- Values 1 and 2 are amounting to 2652 and 2741, respectively.
- Values 3-5 have around 10k+ per each score.

DigitalSkola
PepeThink

Region Category

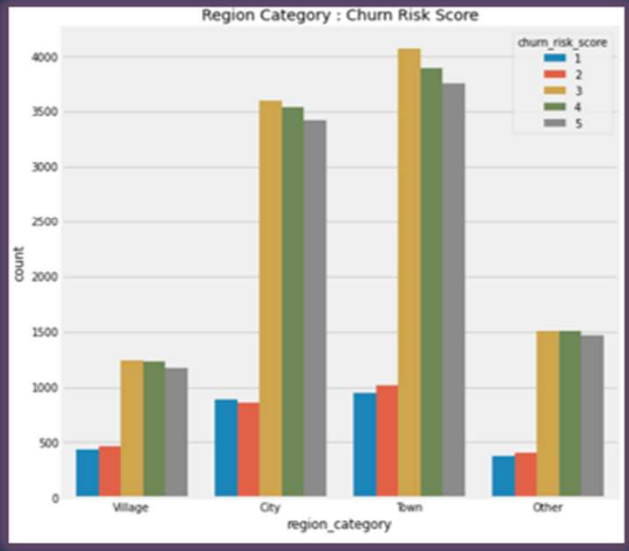


Composition of Region Category

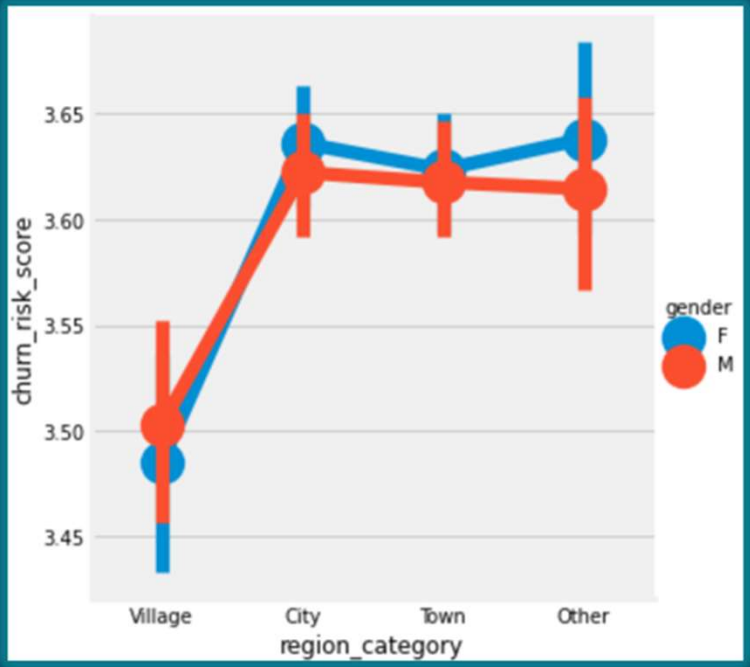
- Because of Null values are 14.69% (quite big portion), they will be grouped under "Others" category.
- Town and City contribute to 72% of the customers' region.
- Village and Others proportion are the remaining 28%.

Interesting Insight:

Most of customers with churn_risk_score 3-5 are from City and Town.



Region Category

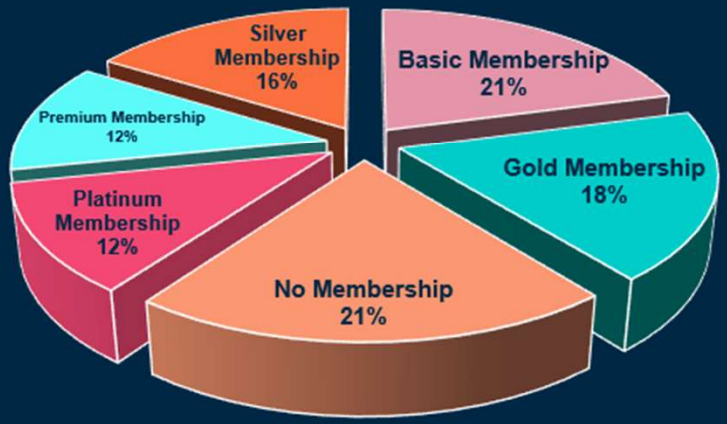


Interesting Insights:

Female customers with higher churn_risk_score from City, Town and Other are higher than Male customers.

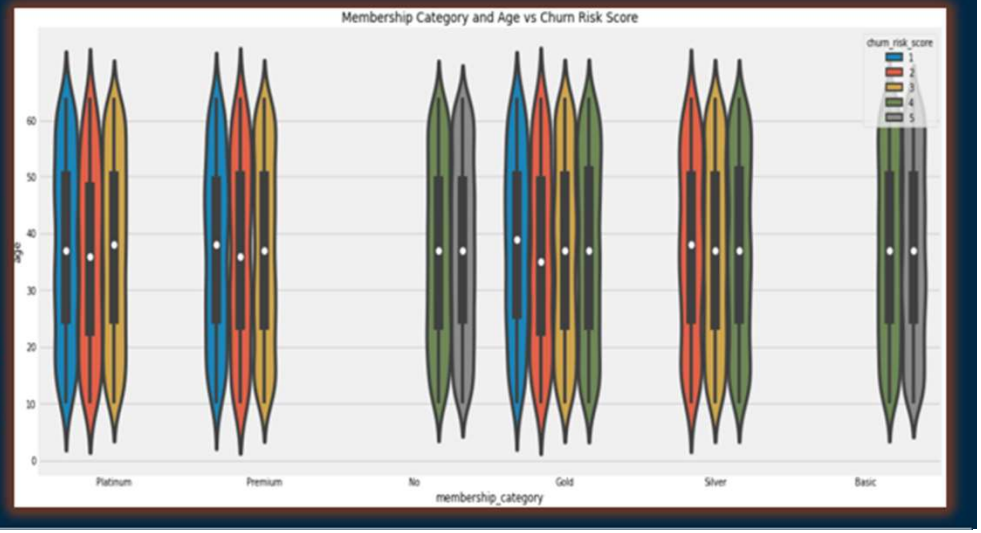
Female customers with higher churn_risk_score from Village is lower than Male customers.

Membership Category



Composition of Membership Category

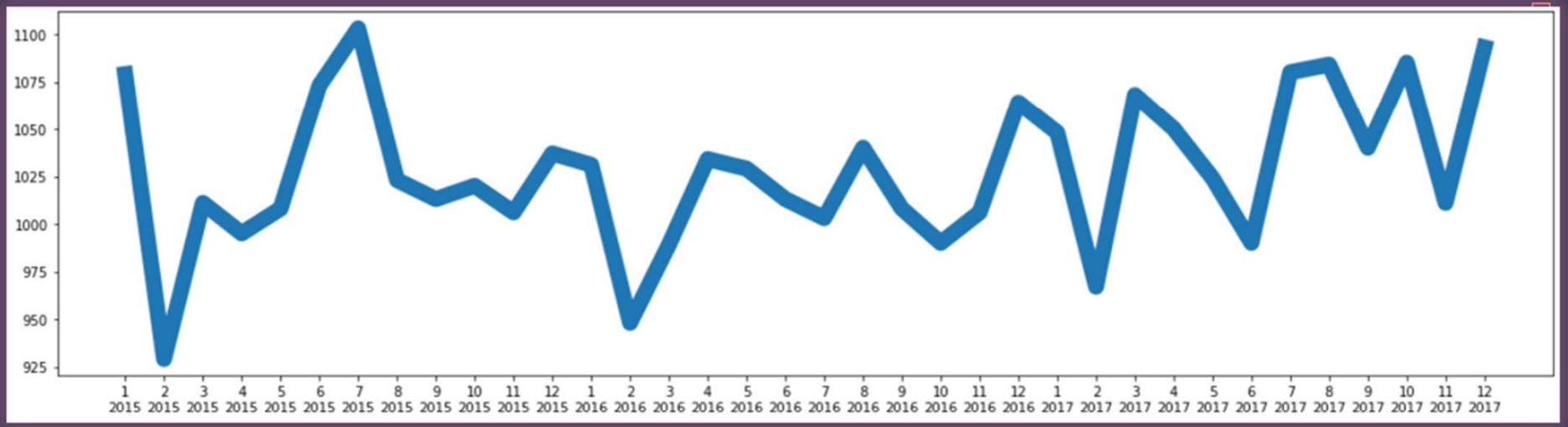
- The proportions for this feature are fairly even.
- The largest proportion is Basic Membership (21%), and the smallest proportion is Premium and Platinum Membership (12%).



Interesting Insight:

Most of customers with churn_risk_score 3-5 has no membership and basic membership.

Joining Date

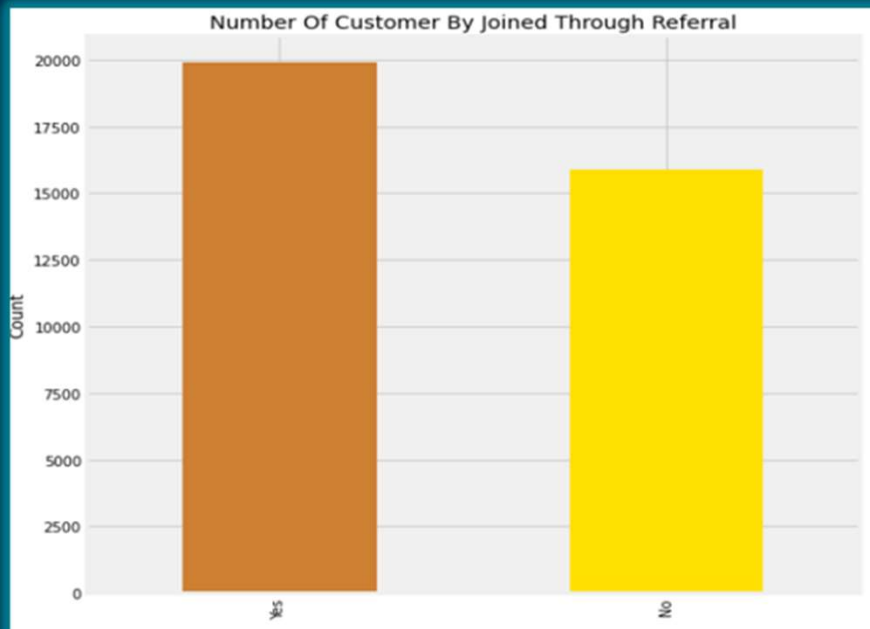


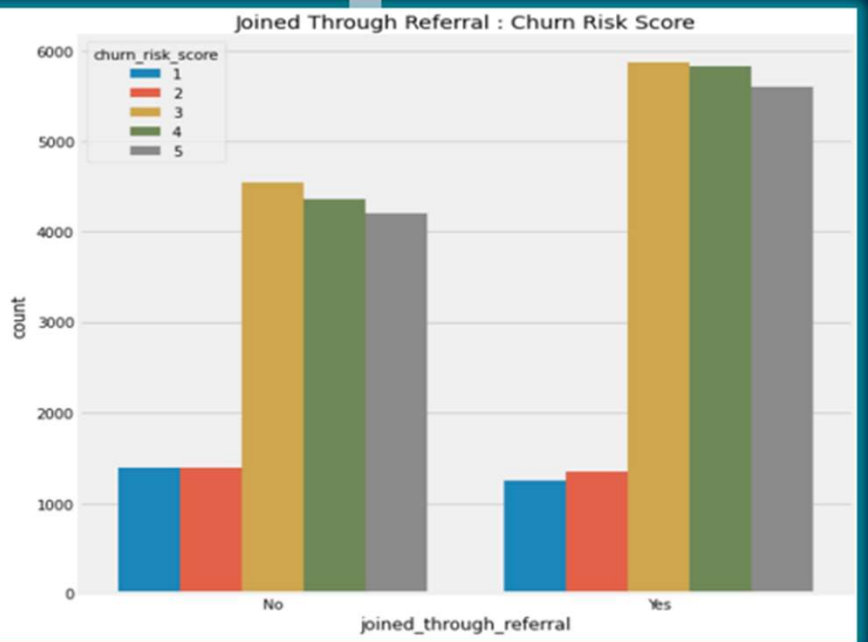
Zoom-IN per month: a Volatile Trend

Zoom-OUT per year: an Up Trend. At least, this is a good signal for the Business.

DigitalSkola
PepeThink

Joined Through Referral





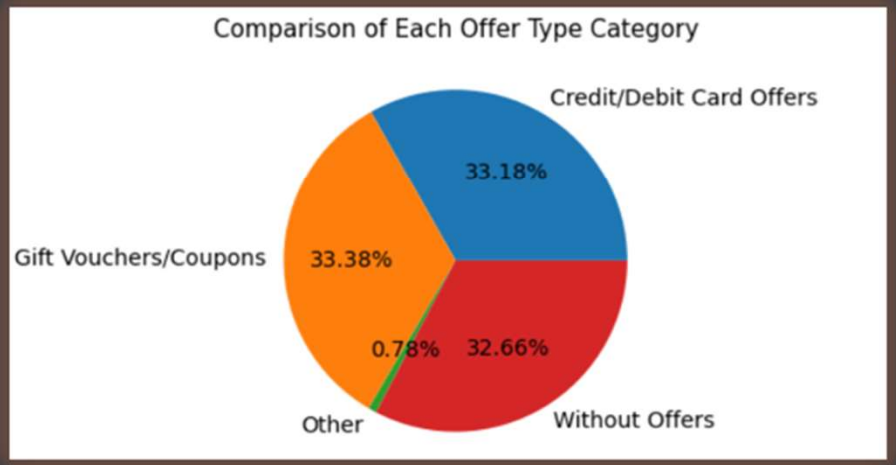
Interesting Insights:

Proportion of customer joined through referral are higher than non referral.

However, churn score of 3-5 are mostly coming from those who are coming from referral program.

Preferred Offer Types

- The three categories have almost the same proportion.
- There is 0.78% of null values. So that null values will be replaced with the mode of the composition (Gift Vouchers/Coupons).

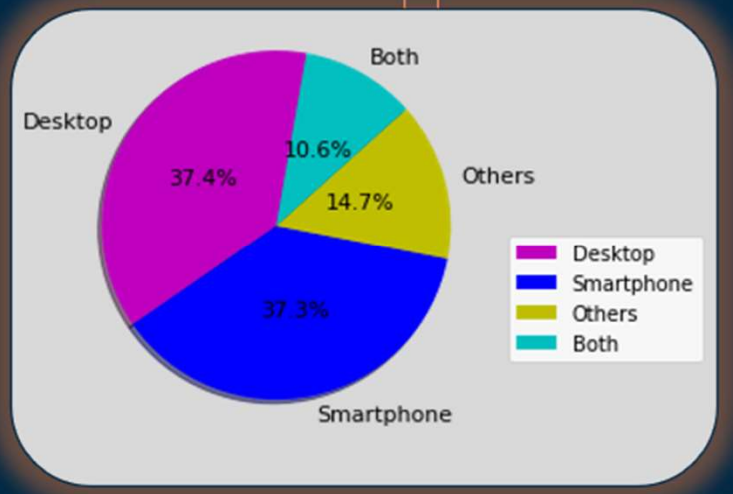


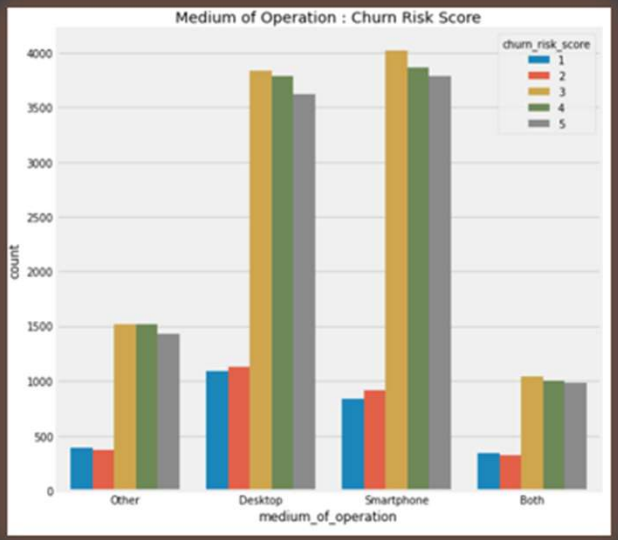
Interesting Insight:

Low churn rate are customers with Gift Vouchers/Coupons.

Medium of Operation

Because of "?" values are 14.7% (quite big portion), they will be grouped under "Others" category.

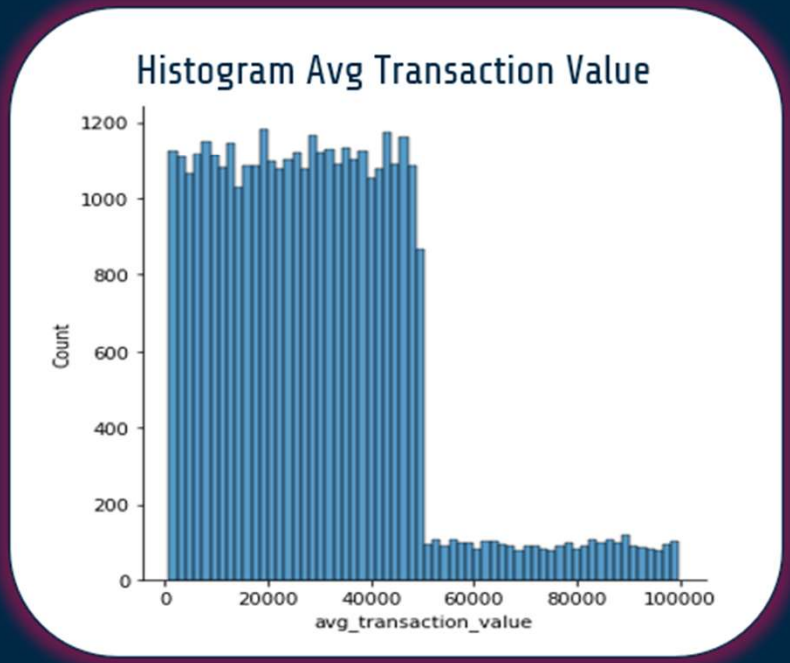


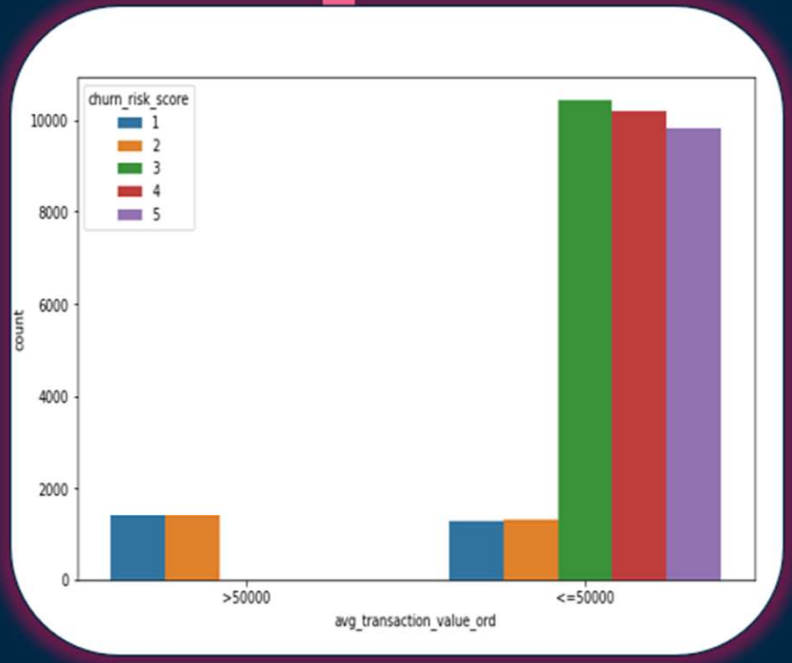


High churn rate are customers who used Smartphone and Dekstop. This is inline with the largest composition of Operation Media.

DigitalSkola
PepeThink

Avg Transaction Value





Interesting Insight:

High churn rate are customers with low average transaction value.

Feedback

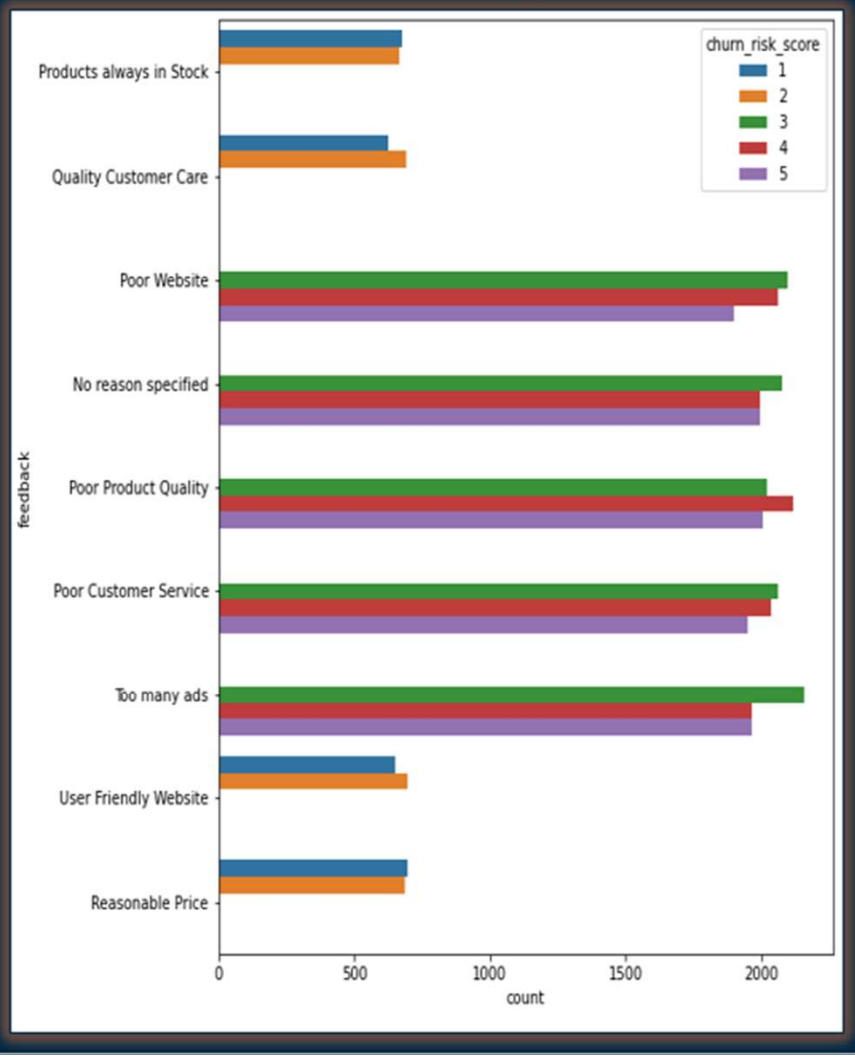
Interesting Insights:

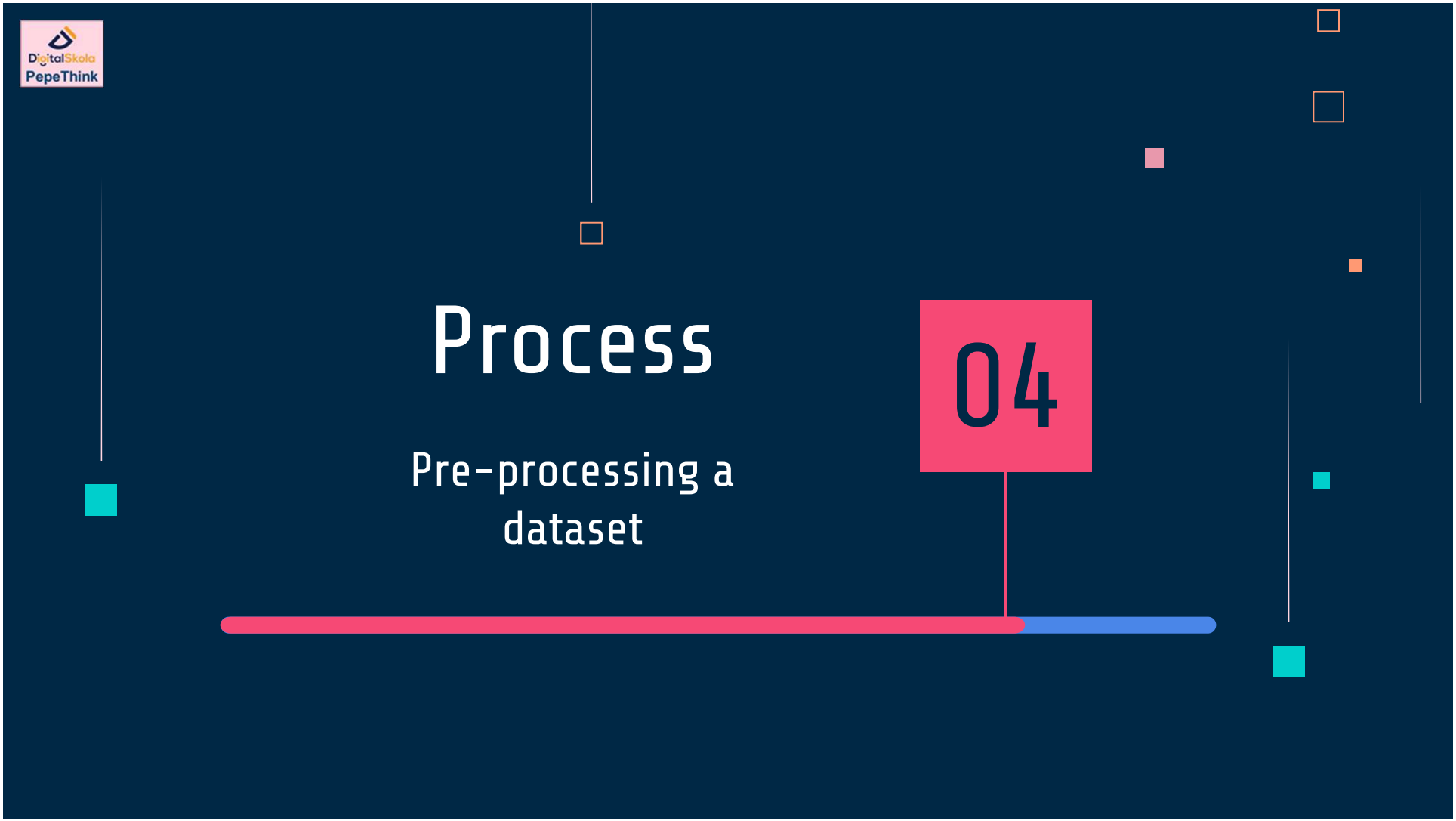
Positive Feedback has a lower churn risk, such as:

Product always in stock, quality customer care, user friendly website, and reasonable price.

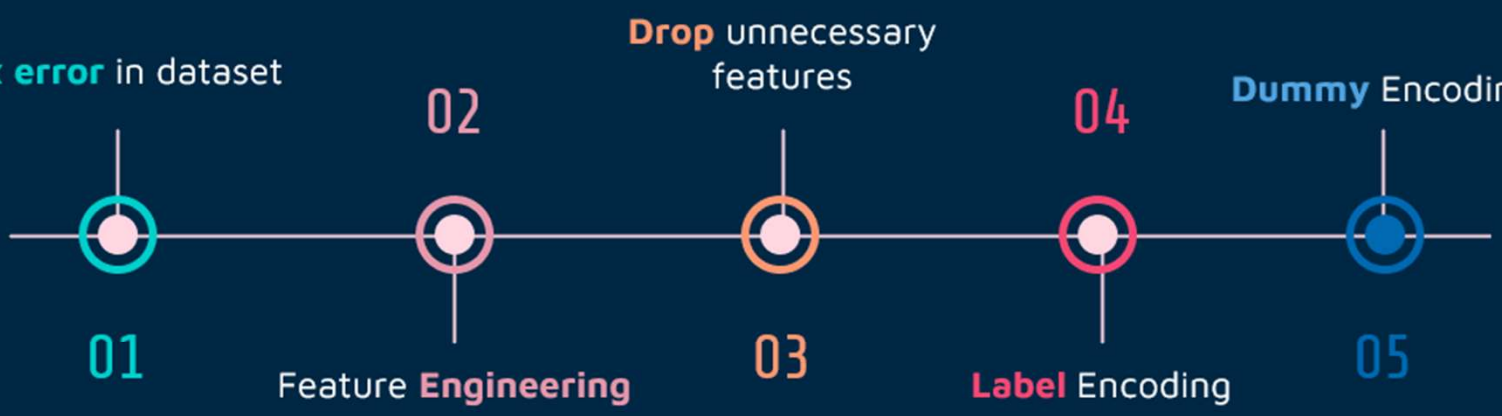
Negative Feedback has a higher churn risk, such as:

Poor website, Poor Product Quality, Poor Customer Service, Too Many Ads, including No Specified Reason.




DigitalSkola
PepeThink
Process
04
Pre-processing a
dataset

Pre-processing



DigitalSkola
PepeThink

01 Fix Error in Dataset

Threshold 5% (based on general best practice)

#	Features	Fixed Item	Pre-processing
1	gender	Unknown value	Below threshold → drop
2	region_category medium_of_operation	Null value "?" value	Above threshold → group "Other"
3	joined_through_referral	"?" value	Above threshold → replaced based on referral_id
4	preferred_offer_types	Null value	Above threshold → replaced with mode
5	avg_frequency_login_days points_in_wallet	Error value Null value	Above threshold → replaced with mean

DigitalSkola
PepeThink

02 Feature Engineering

joining_year	last_visit_hours
Grab only four-digit-year	Grab only two-digit-hour

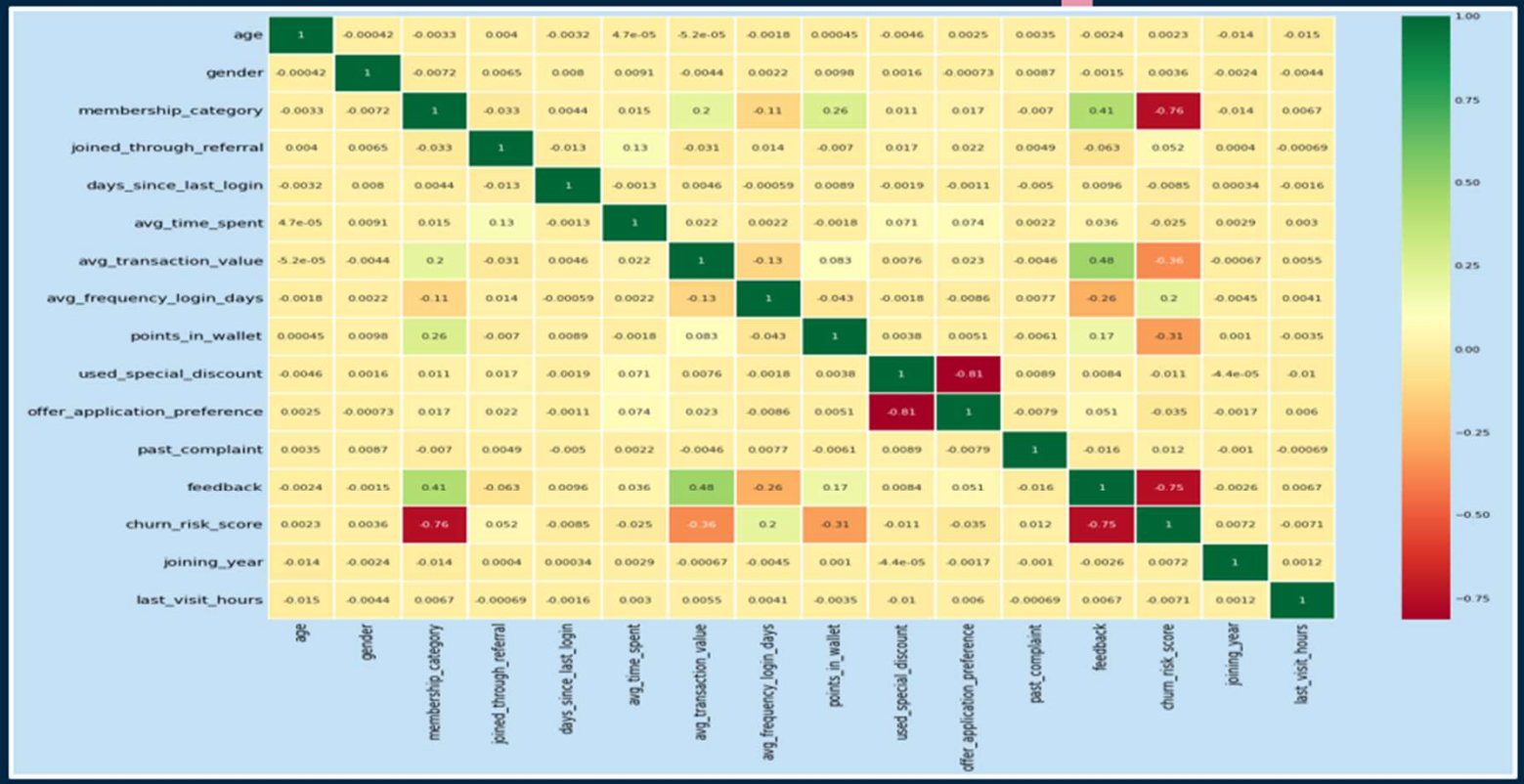
03 Dropping Unneeded Features

Dropping unneeded features based on Exploratory Data Analysis:

Categorical
customer_id
Name
security_no
referral_id

Numerical
joining_date
last_visit_time

Correlation Matrix



04 Label Encoding

code	membership	feedback	join_year	gender	join_through_referral	special_discount	offer_preference	past_complaint
0		2015	Poor Product Quality, No reason specified, Poor Customer Service, Poor Website, Too many ads	♂				
1		2016	Quality Customer Care, User Friendly Website, Products always in stock, Reasonable Price	♀				
2		2017						
3								
4								
5								

05 Dummy Encoding

complaint_status	region	medium_of_operation	preffered_offer	internet_option
	Village	Other	Without Offers	Wi-Fi
	Town	Smart Phone 		Mobile
	City	 Desktop		Fiber_Optic
	Other	Both 		
				

Develop

05

Develop model with evaluation and recommendation

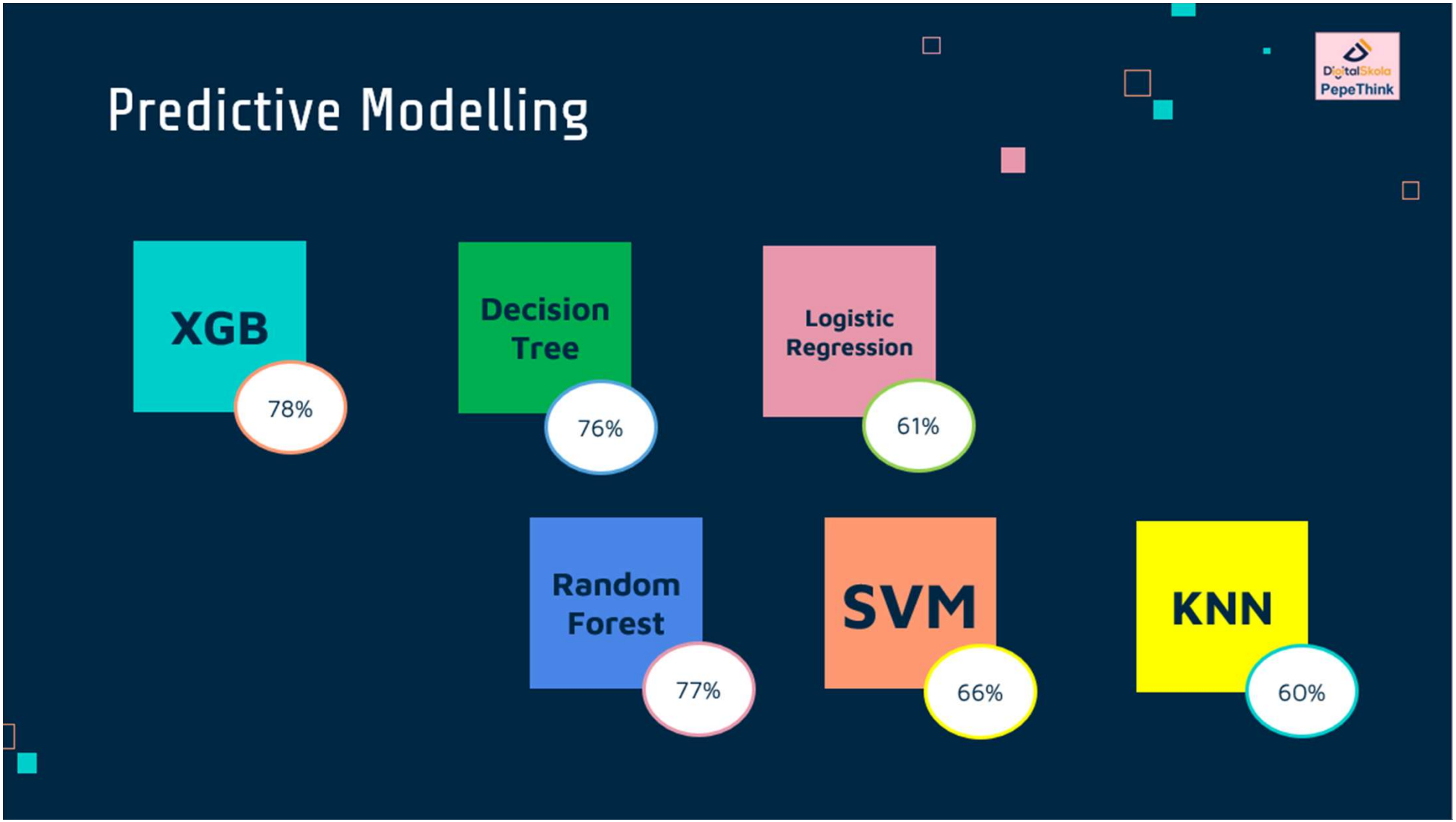
Feature Scaling

We normalized the range of data features using Scaler.

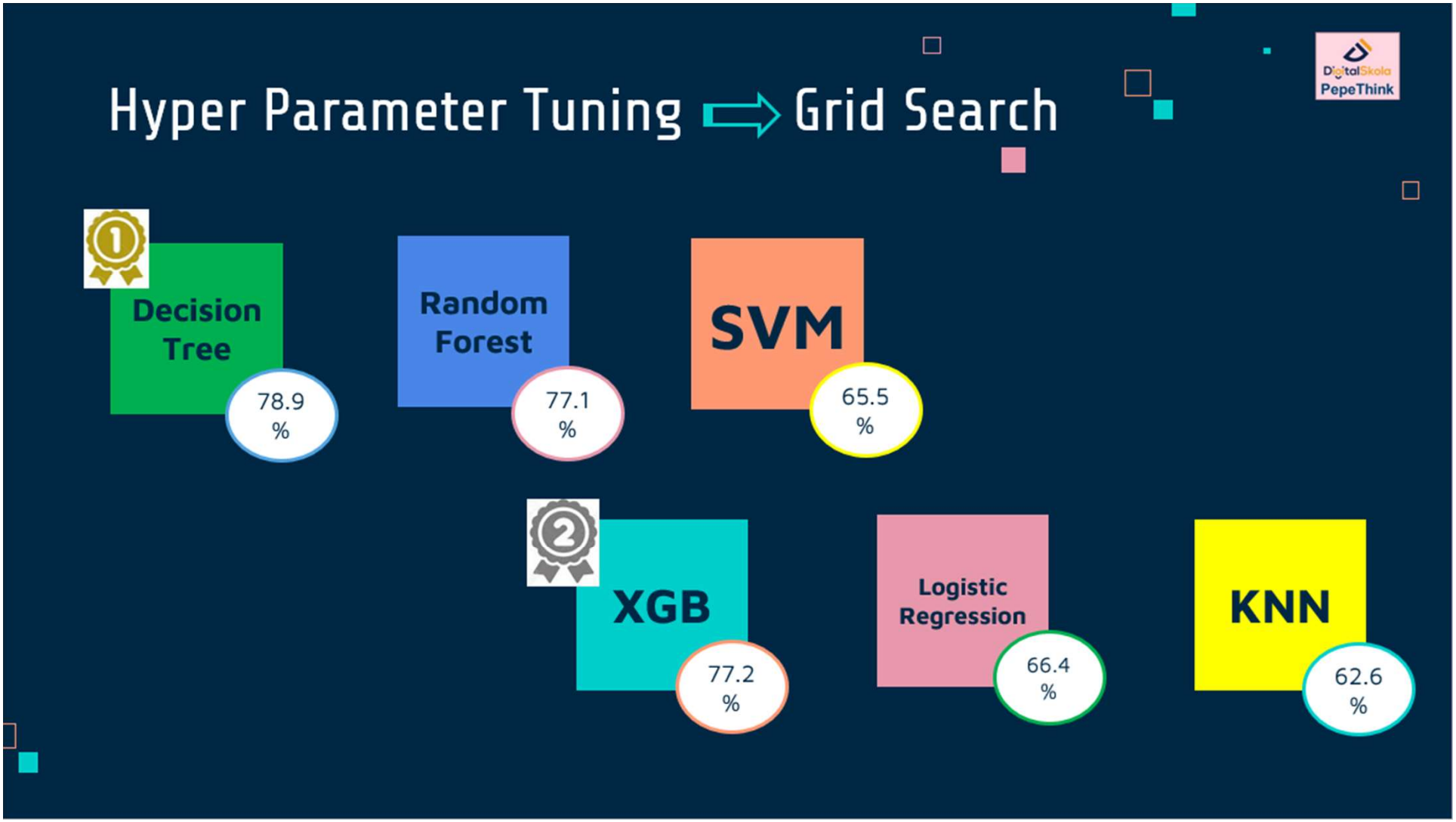
SMOTE

With Synthetic Minority Oversampling Technique (SMOTE), the imbalance classification in the dataset can be catered.

Predictive Modelling



Hyper Parameter Tuning ⇨ Grid Search



Grid Search Result

```
Decision Trees Test Accuracy: 0.7888190076869322
Decision Trees Best Params: {'DT__criterion': 'gini', 'DT__max_depth': 6, 'DT__min_samples_leaf': 5, 'DT__min_samples_split': 2}
```

```
XGBoost Test Accuracy: 0.7724668064290706
XGBoost Best Params: {'XGB__learning_rate': 0.3, 'XGB__max_depth': 6, 'XGB__min_child_weight': 1, 'XGB__n_estimators': 150, 'XGB
  subsample': 0.5}
```

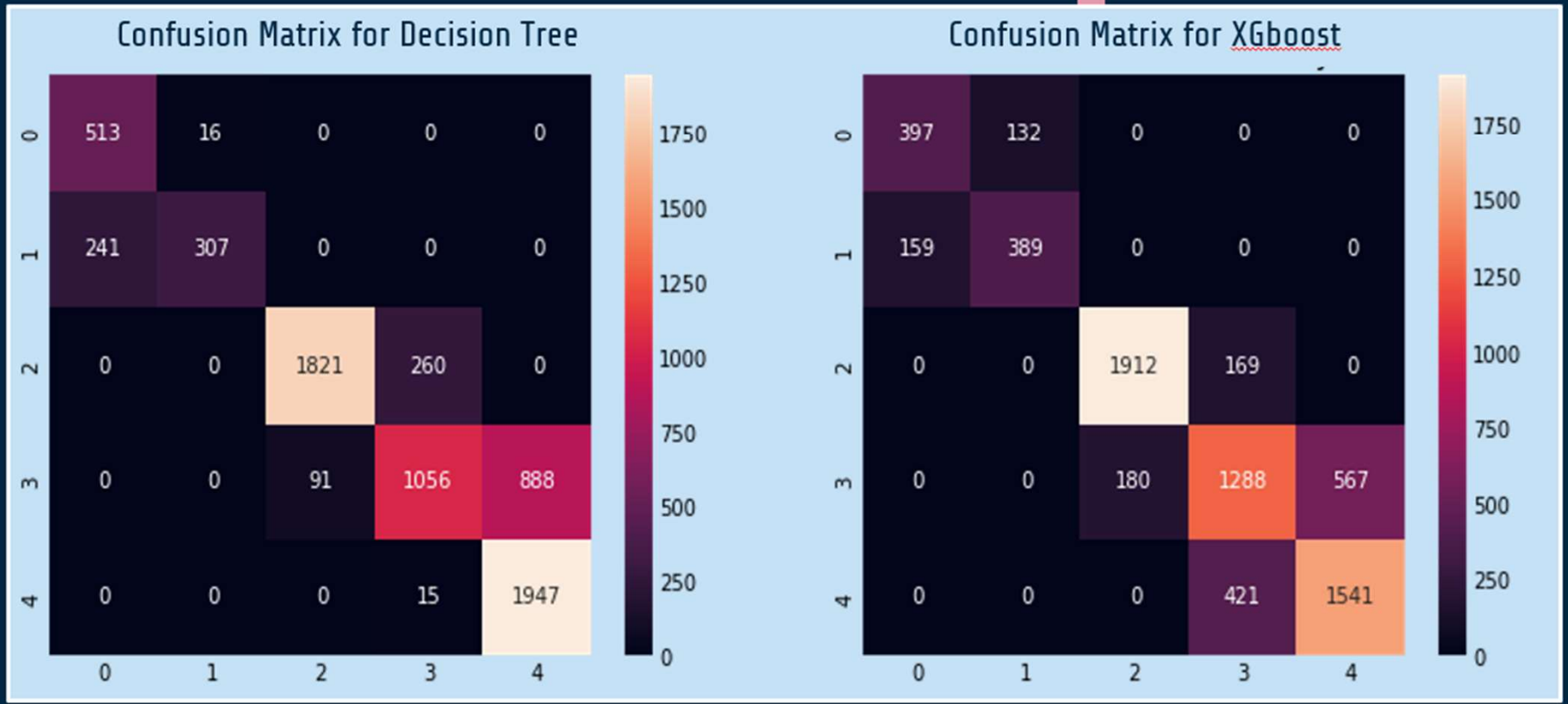
```
Random Forest Test Accuracy: 0.7710691823899372
Random Forest Best Params: {'RF__max_depth': 6, 'RF__min_samples_leaf': 3, 'RF__min_samples_split': 2}
```

```
Logistic Regression Test Accuracy: 0.6642907058001397
Logistic Regression Best Params: {'LR__C': 0.1, 'LR__penalty': 'l2', 'LR__solver': 'liblinear'}
```

```
Support Vector Machines Test Accuracy: 0.6545073375262055
Support Vector Machines Best Params: {'SVM__C': 6, 'SVM__kernel': 'rbf'}
```

```
K-Nearest Neighbors Test Accuracy: 0.6261355695317959
K-Nearest Neighbors Best Params: {'KNN__metric': 'manhattan', 'KNN__n_neighbors': 6, 'KNN__weights': 'distance'}
```

Confusion Matrix



Develop Model

We recommend to use **Decision Tree** and/or **XGB** as the model because of the highest accuration score compared to other algorithms (78.9% and 77.2%, respectively).

Evaluation

Useful **insights** from data to know our customer better:

1. Customer that **loyal** to the Company are providing positive feedback, lots of transactions and utilize the Gift Vouchers/ Coupons.
2. Customer that tend to **churn** are leaving negative comments, join through referral program, have no/only basic membership, located in City and Town with female customers are more disappointed than male.

DigitalSkola
PepeThink

Recommendations

1. **Marketing team** to engage with more merchants for Gift Vouchers/ Coupons. In addition, review the referral program and consider to assign Relationship Manager for each new customer (provide all information required and other benefit of paid membership). Furthermore, special program for City and Town customers can be prioritized to attract female customers.

2. **Product team** to improve our product quality. Focus can be concentrated to improvement of existing product instead of new product launch. Product stock should also be always monitored.

3. **Customer Service team** to give Service Excellent to customers. Also, need to set out standard of Service Level Agreement that follow industry practice, response in media social is recommended no later than 48 hours.

4. **Digital team** to revamp Company's website and manage the frequency of Advertisements interfacing.

DigitalSkola
PepeThink

THANKS

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon, and infographics & images by Freepik