ADVENTURE HARDWARE GROUP DATA SCIENCE PROJECT

ADVENTURE HARDWARE GROUP

- Business Objective
- Reporting Dashboard
- Recommendation Engine
- Predicative Model
- Recommendations
- References

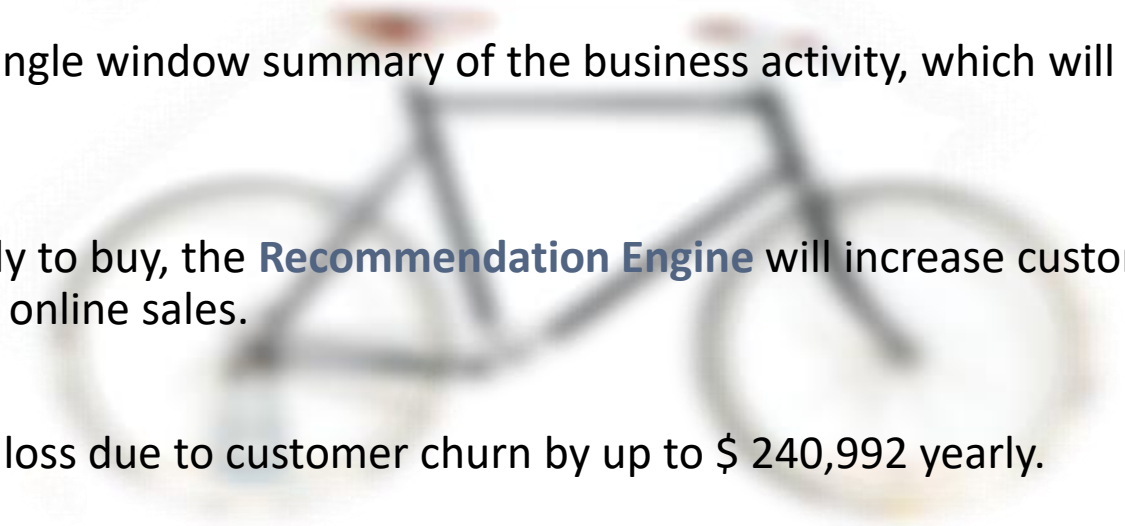
About AHG…

- Adventure Hardware Group (AHG) was established in 2011 and is a global manufacturing organization with operations in America, Europe . AHG operates in 6 Regions and the main products are Bikes, Accessories, Clothing and Components. These products are sold using online Channel using web and Reseller Channel delivering services through stores.

- AHG has recently engaged Kernel Decision Science Limited to help it develop a data architecture that will enable it to leverage all data assets across the organisation and regions. More recently, AHG has developed an ambition to leverage data from weather, climate change, Google search trends, Twitter and Facebook to understand correlations between business decisions, external forces and company performance. Kernel recommends that an integration of all traditional and digital data sources will deliver higher commercial value and strategic opportunities to help AHG take new markets and develop better and more relevant products for existing and new customers.

Business Objectives

- To develop an automated report system that will be creating a stream reporting solution that will deliver analytics.
- The 3 tools built in this project will empower the management of AHG to make data driven decisions, and in-turn increase the commercial value of the business.
- The **Reporting Dashboard** will enable a single window summary of the business activity, which will significantly reduce the decision making time.
- By suggesting products a customer is likely to buy, the **Recommendation Engine** will increase customer basket sizes and the revenue generated through online sales.
- The **Predictive Model** is estimated to cut loss due to customer churn by up to $ 240,992 yearly.



Business Objectives/Requirements

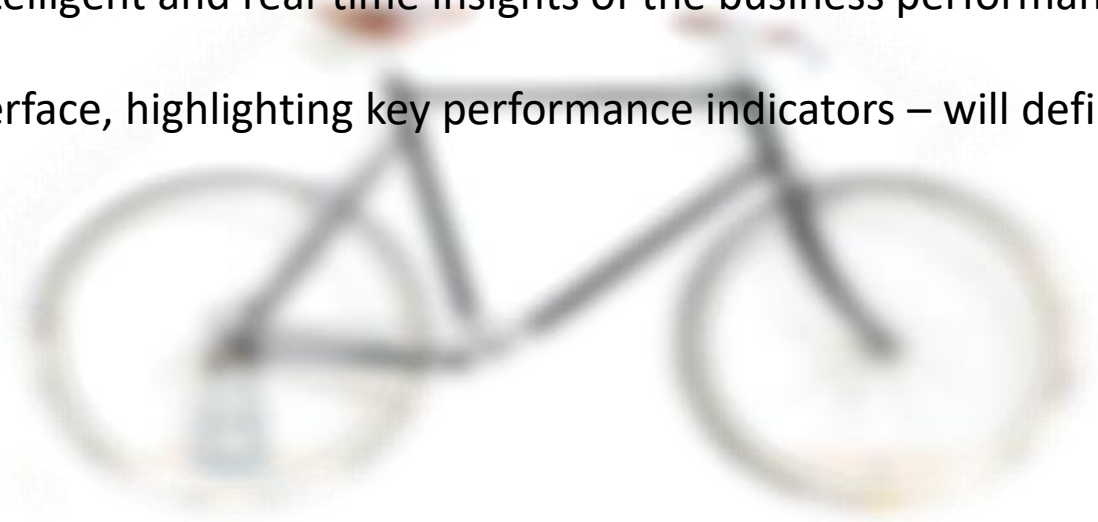
- To develop an automated report system that will be creating a stream reporting solution that will deliver analytics.

Reporting Dashboard	• Need for Reporting Dashboard • Key Performance Indicators • Dashboard Visuals
Recommendation Engine	• Why recommendation engine • Methodology • Data processing, model training, evaluation and selection • Model deployment
Predictive Model	• Purpose of a predictive model • Data exploration • Model training, evaluation and selection • Commercial impact of churn • Recommendations to prevent churn

Reporting Dashboard

Reporting Dashboard -Purpose

- As a global manufacturer and seller of bikes and bike accessories AHG continuously generates enormous data everyday - across 6 countries and 4 product categories.
- The management at AHG need – visual, intelligent and real-time insights of the business performance.
- A dashboard with a clean and intuitive interface, highlighting key performance indicators – will define and support AHGs business decisions.



Reporting Dashboard -KPI

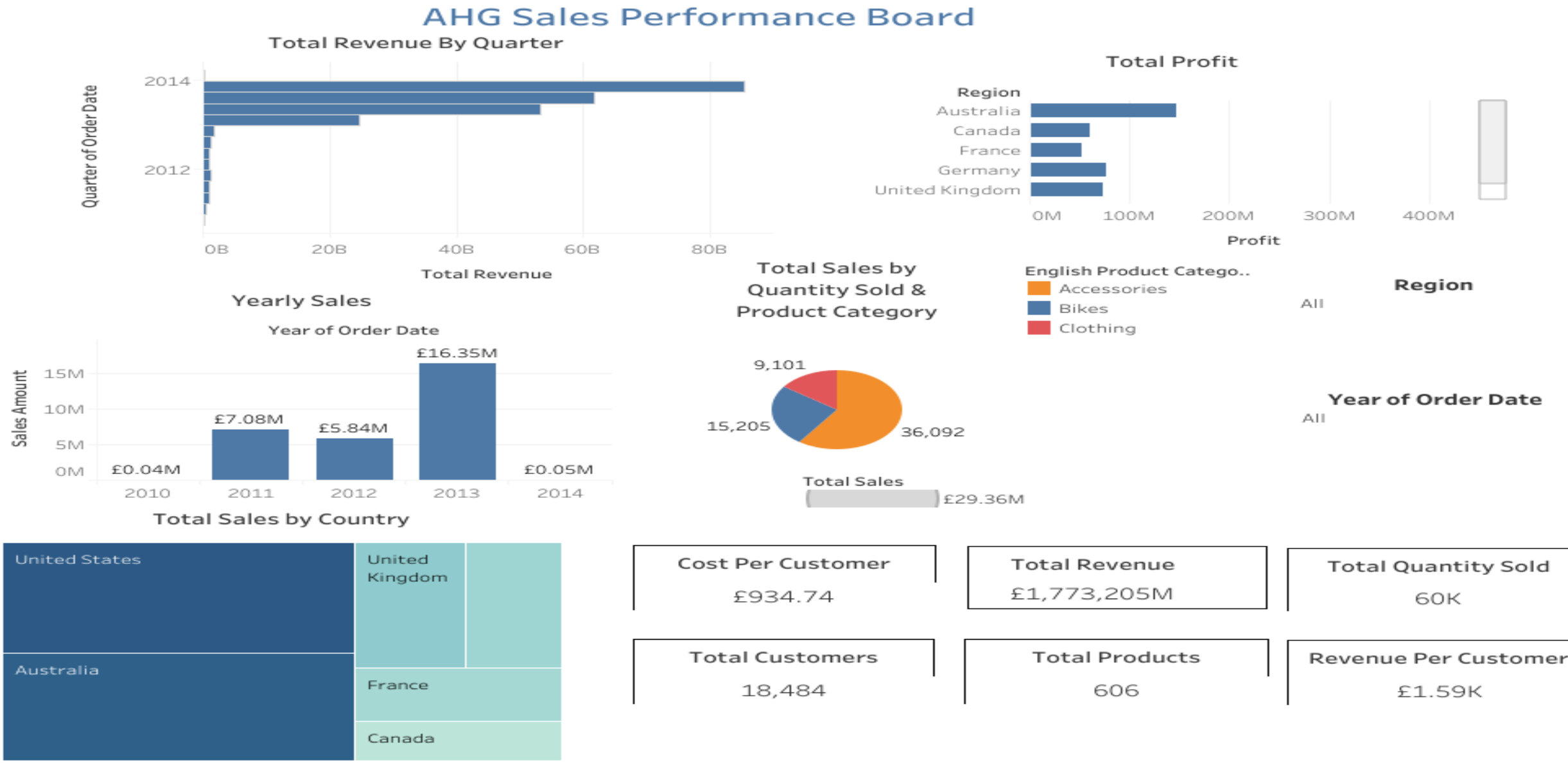
Key Performance Indicators

- Net Revenue
- Total Product Cost
- Total Profit
- Total Customers

- Total Revenue by Country & Region
- Net Revenue by Quarter
- Total Transaction
- Total Item Sold

- Profit by Product Category
- Revenue & Order Quantity by Product Category

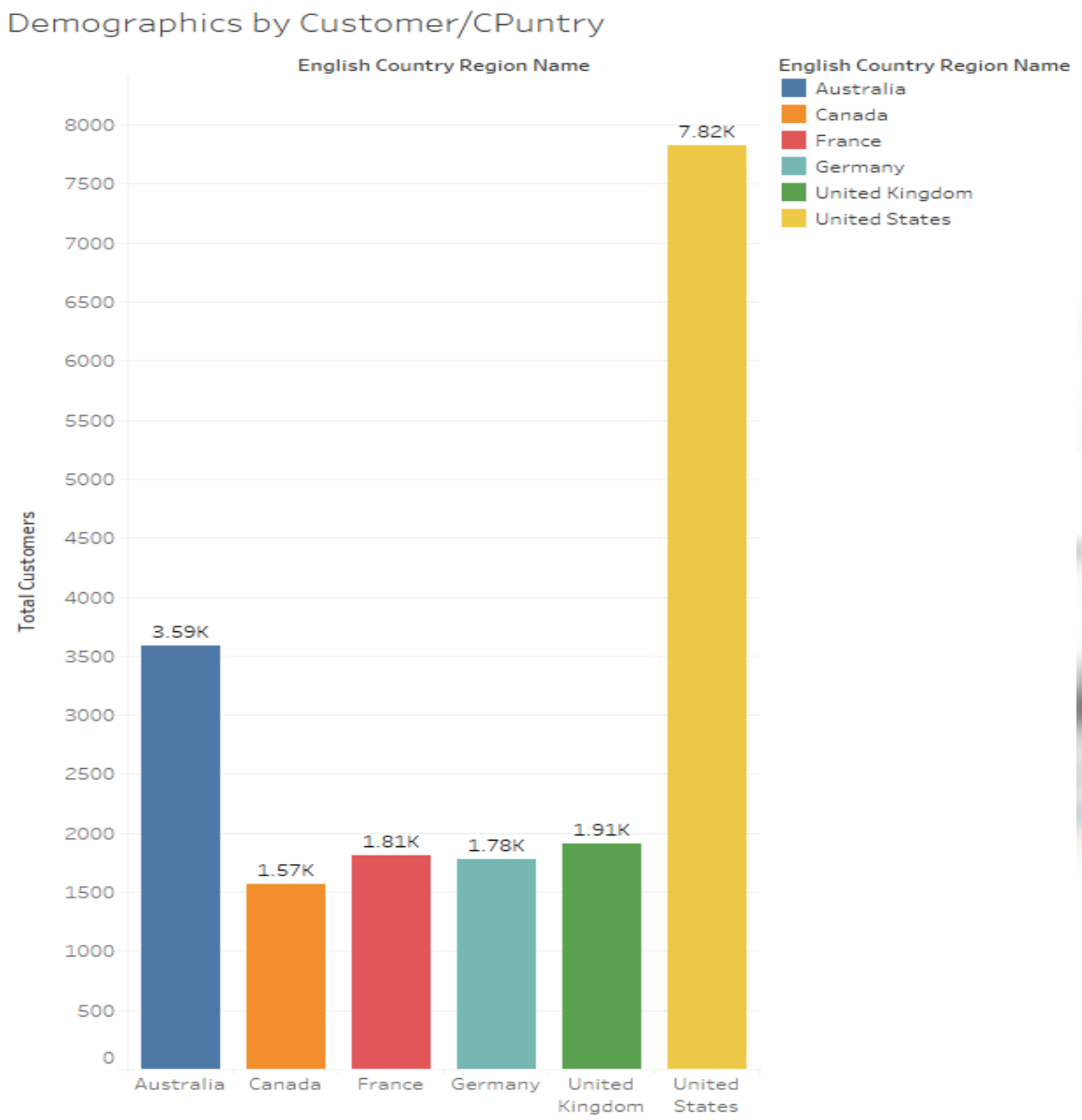
Tableau Dashboard



AHG Dashboard/Demographics by Occupation

- The highest Total Sales are generated by US with £5.64M in 2013.
- The highest sales is mainly generated with occupation in Clerical ,Skilled Manual.
- The Australia is second largest with £4.34M in 2013.
- The sales of £4.34 M in Clerical, Manual and Skilled Manual.

Demographics by Customer/Country



- The customer demographics from US has the highest AHG product buyers.-7.82k
- The income demographics from Australia are the AHG product buyers.-3.59k

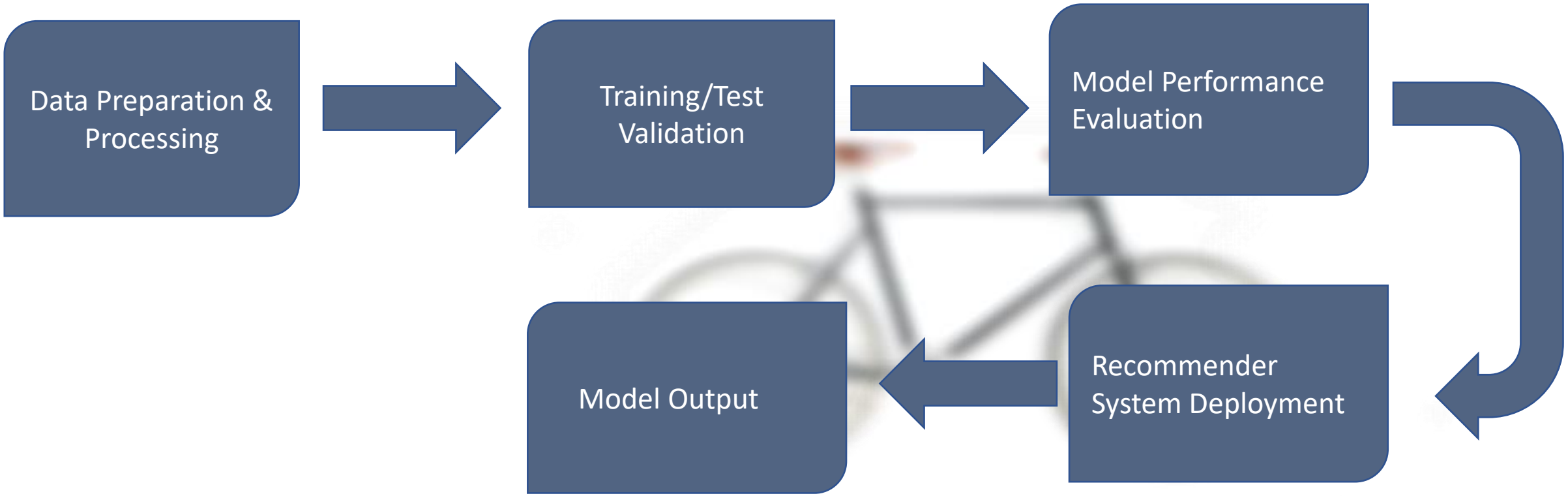
Recommendation Engine

Recommendation Engines are an excellent way for brands to keep customers engaged by recommending relevant content to increase online sales

Need for Recommendation Engine

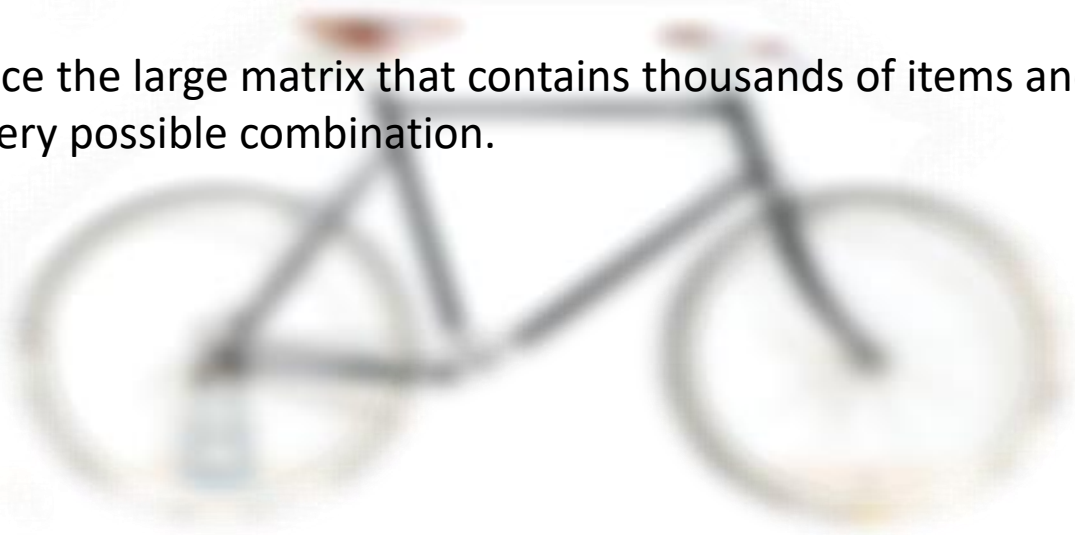
- The key focus for the need of recommendation engine is to provide AHG with a solution that will automatically recommend a product to a customer based on their purchase history.
- AHG has been selling bikes online over the last 4 years. Last year they introduced clothes and accessories as new categories online
- Online sales have contributed to 26.8% of Total sales with a profit of 118% Total profit
- Recommendation engines are an excellent way for brands to keep customers engaged by recommending relevant content to increase online sales
- This will retain the customers based on purchase history and increase the commercial value of the business for online sales channel.

Methodology used for Recommendation Engine



Data Processing

- The Dataset was processed to remove any missing values thereby matching all the purchases to a specific customer.
- A lookup table was created to keep track of each Product ID a description of the item.
- The Product ID and Customer ID were grouped together thus creating grouped purchased quantities to replace a sum of zero purchase with one to indicate purchased item.
- The sparse ratings matrix was set up to reduce the large matrix that contains thousands of items and thousands of users with a user/item value required for every possible combination.



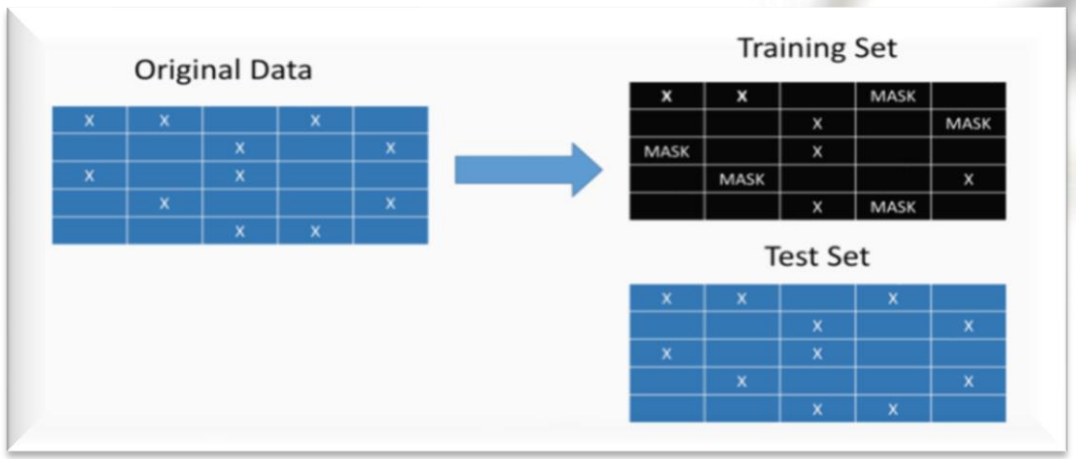
Data Training and Validation Test

- Using the Random library, both the training and test set were made the exact copy of the original dataset. However, a random percentage of user/item interaction were masked by the training set to act as if the user never purchased the item. i.e. making it sparse entry with zero.

- The splitting of the Dataset returned the training set a test set that has been binarized to 0/1 for purchased/not purchased

- A list of which users had at least one item masked.

- In this project, 20% of the user/item interactions were masked



Algorithm Implementation

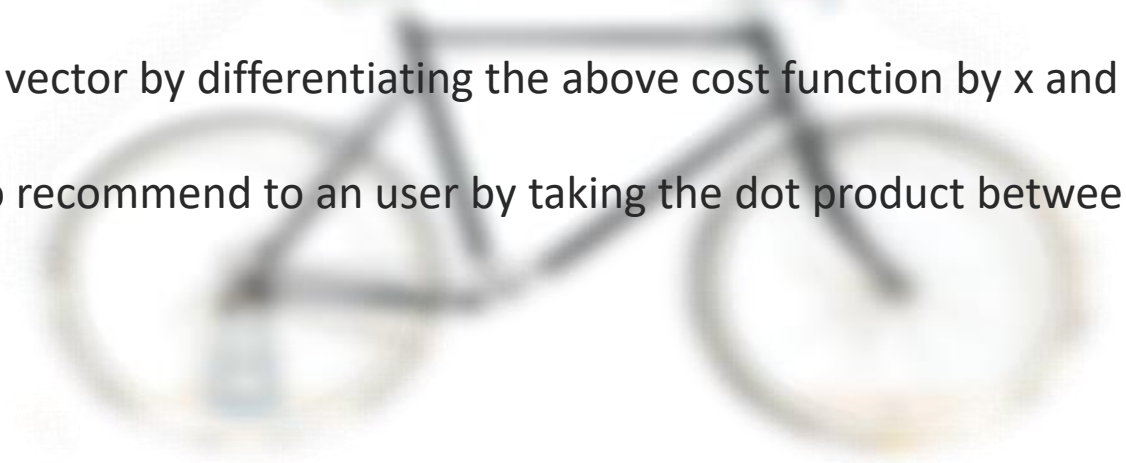
- Alternating Least Squares: It is a matrix factorization algorithm that was used in this item based collaborative filtering to turn our sparse rating matrix into a confidence matrix.

- Where Cui is the confidence matrix for our users u and item i. The α term represents a linear scaling of number of purchases and the rui is the original matrix of purchases

- The cost function was followed to find the User interaction and Item interaction matrices.

- We find the user(x) vector and the item(y) vector by differentiating the above cost function by x and y respectively.

- We find the largest p value having items to recommend to an user by taking the dot product between the user and item vectors(U and V).



Model Performance and Evaluation

Using the metric library

- The ROC was defined with a function using the 20% of the training set that had the purchases masked.

- An AUC of 0.77 was computed after running the algorithm.

- It showed that the system was recommending items the user in fact had purchased in the test set far more frequently than items the user never ended up purchasing.

Deploying the Recommender System

RECOMMENDATION ENGINE COMMERCIAL IMPACT

- Total Number of Customers – 18484
- Number of Items recommended per customer – 10
- Commercial Benefit (based on Standard Cost) if all customers buy one of their top 10 recommended is £38k
- Commercial Benefit (based on List Price) if all customers buy one of their top 10 recommended is £38k

Predictive Model

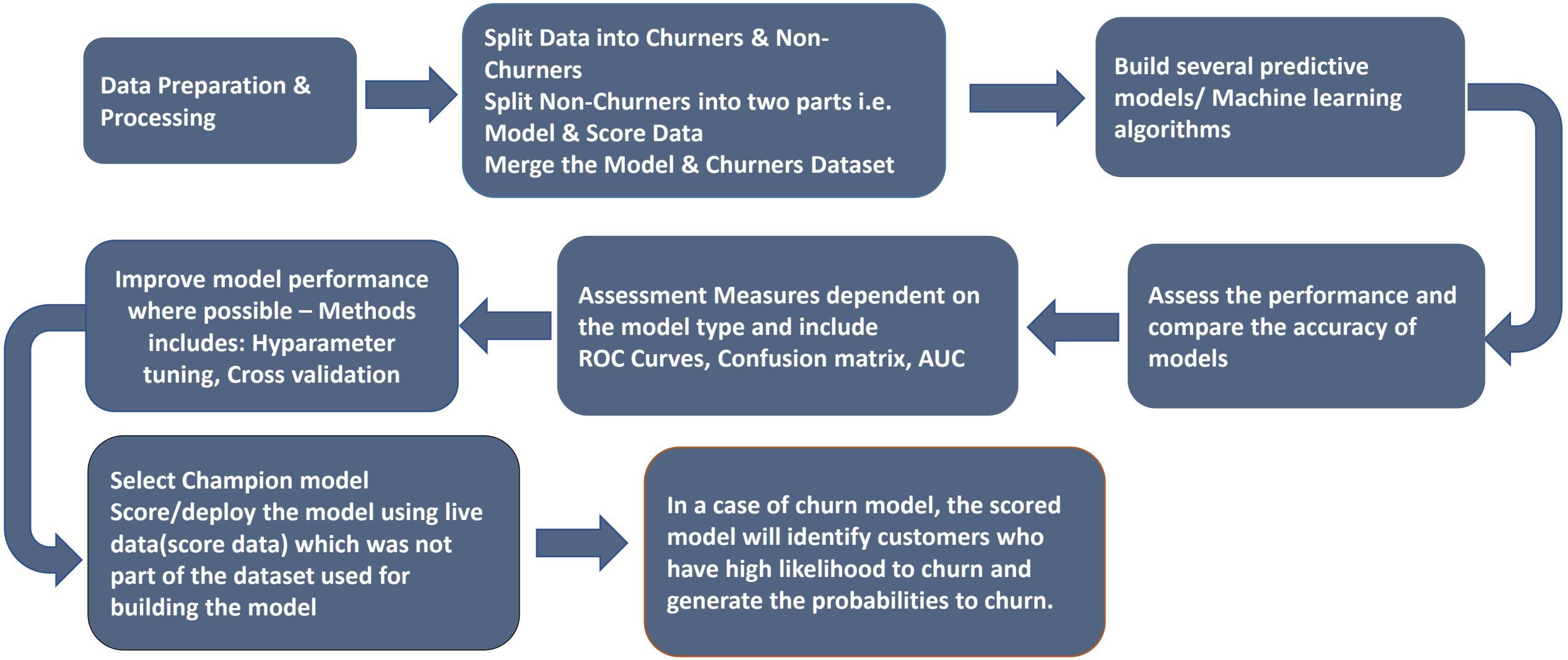
Predictive Model

- It is imperative for AHG as a business entity to be proactive and pragmatic in staying afloat in the business sector. And one of the ways of being ahead of their competitors in the industry is by identifying and understanding their customers with regards to meeting their expectations and retaining them.

- Predictive model is that process that will help AHG in predicting the customers that have the potential of leaving them for their competitors, a term known as “customer churn” which one of the biggest expenditures of an organization. If AHG could figure out why a customer leaves and when they leave with reasonable accuracy, it would immensely help them to strategize their retention initiatives manifold.

- Out of 18,484 customers from AHG database, 5,567 customers have churned having stopped buying from them in the last 8 months.

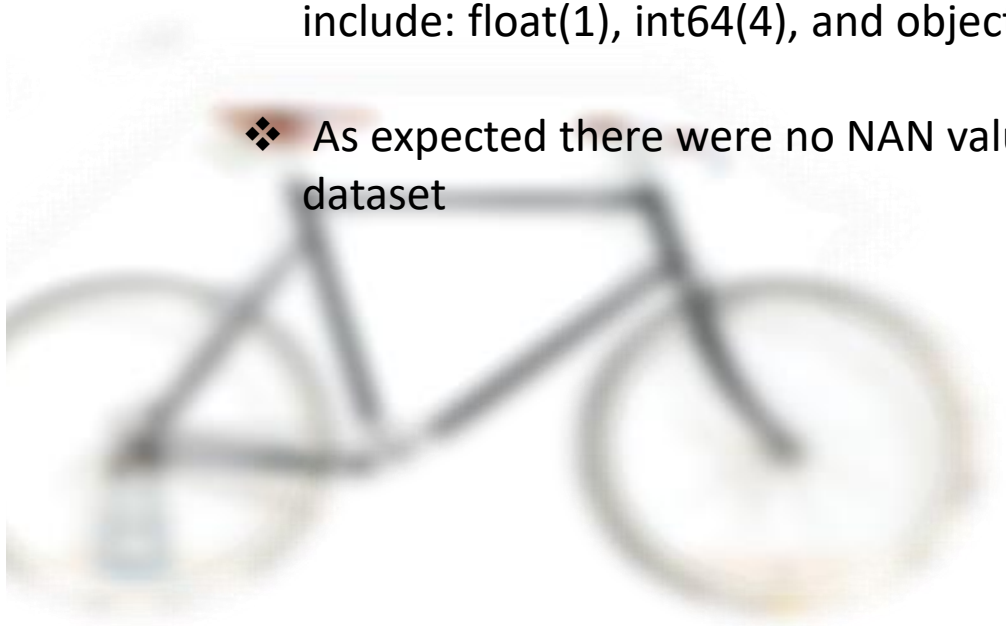
Predictive Model Probable Solution



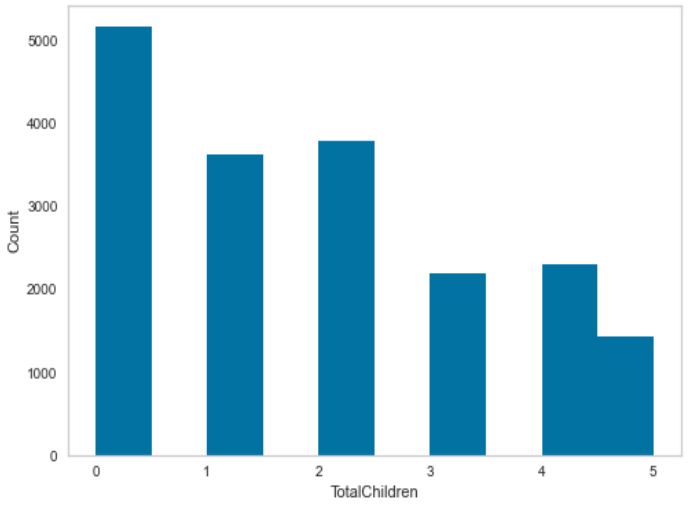
Exploratory Data Analysis

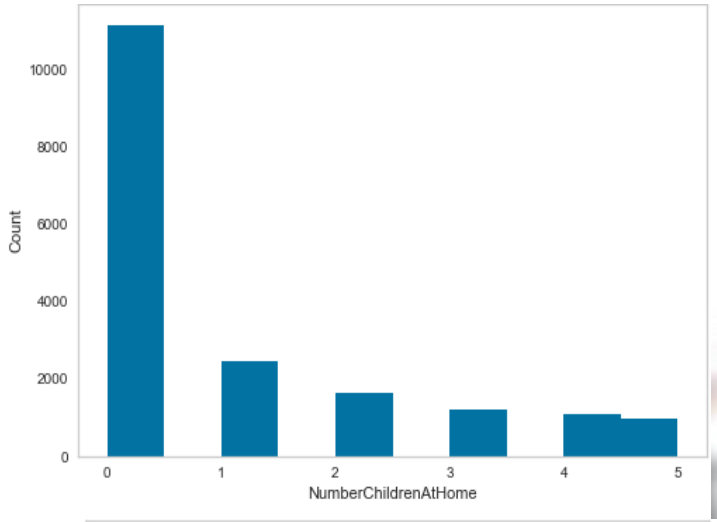
```
RangeIndex: 18484 entries, 0 to 18483
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CustomerID            18484 non-null  int64
 1   BirthDate             18484 non-null  object
 2   Marital_Status        18484 non-null  object
 3   Gender                18484 non-null  object
 4   YearlyIncome          18484 non-null  float64
 5   TotalChildren         18484 non-null  int64
 6   NumberChildrenAtHome  18484 non-null  int64
 7   EnglishEducation      18484 non-null  object
 8   EnglishOccupation     18484 non-null  object
 9   House_Ownership       18484 non-null  object
 10  Car_Ownerdhip         18484 non-null  int64
 11  CommuteDistance       18484 non-null  object
dtypes: float64(1), int64(4), object(7)
memory usage: 1.7+ MB
```

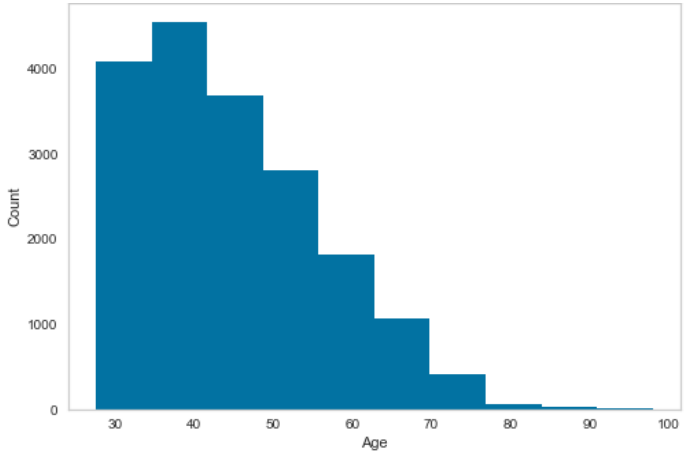
- The Demographic dataset used to plot the distribution of the demographical variables has 12 features and made up of the following datatypes to include: float(1), int64(4), and object(7) .
- As expected there were no NAN values in the dataset

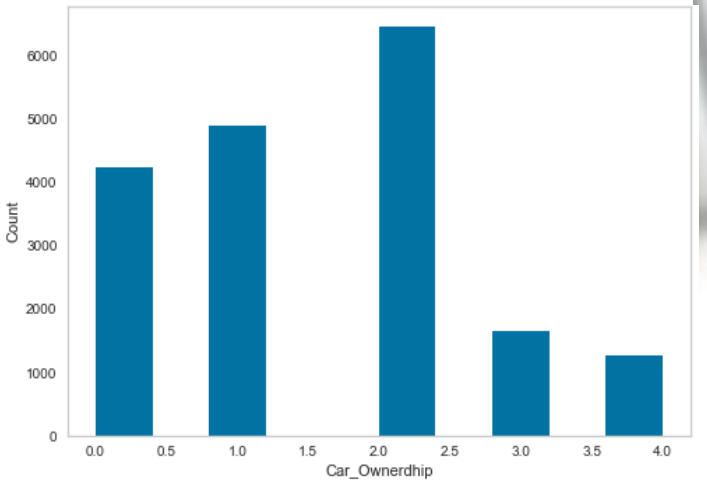


Exploratory Data Analysis





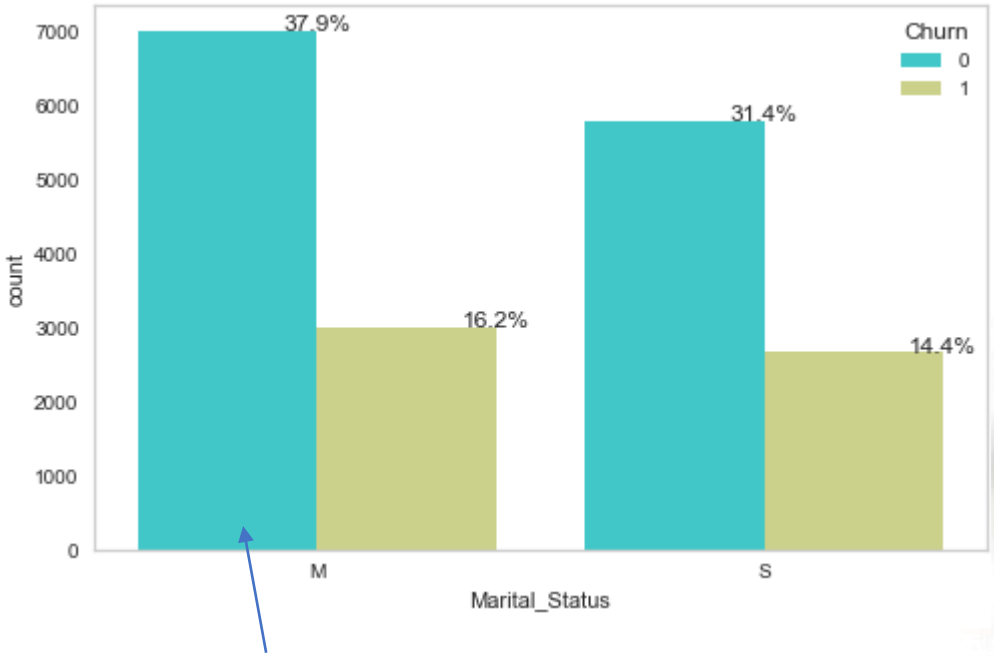




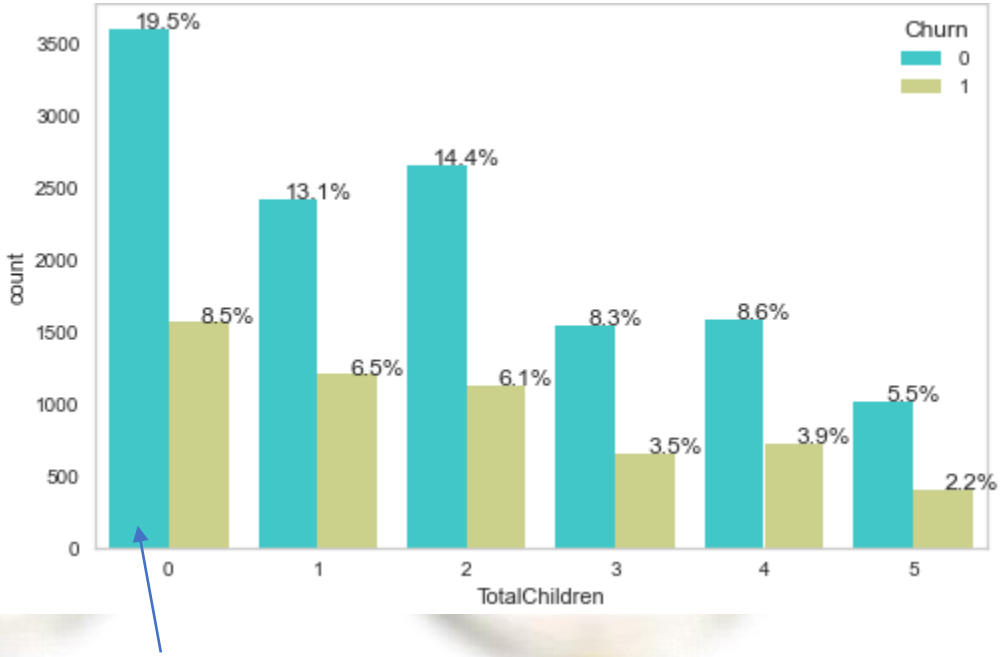
The distribution of continuous variables of the demographic dataset, we confirm that they are normally distributed, and that there are no outliers

Exploratory Data Analysis

Upon plotting the input variables versus the target variable churn, we are able to visually identify certain key characteristics of the customers



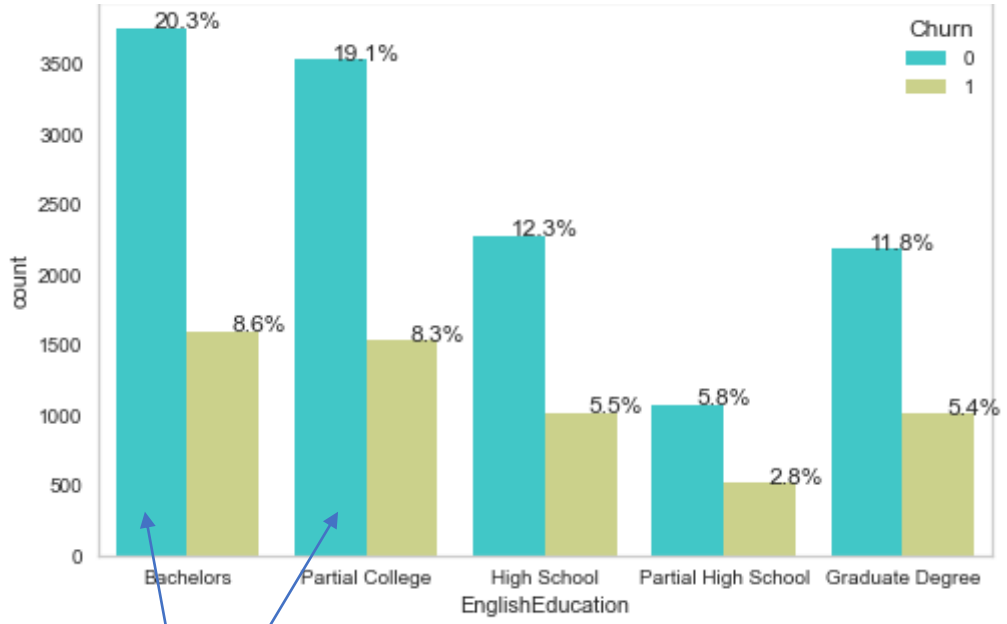
Married customers are less likely to churn than single customers

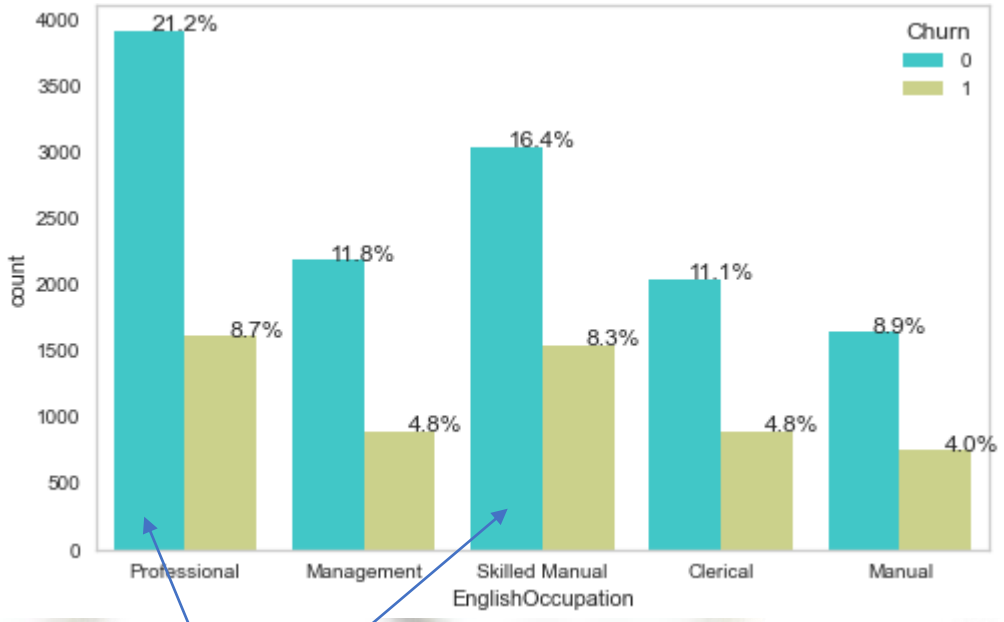


A significant number of customers without children are non churners

Exploratory Data Analysis

Upon plotting the input variables versus the target variable churn, we are able to visually identify certain key characteristics of the customers



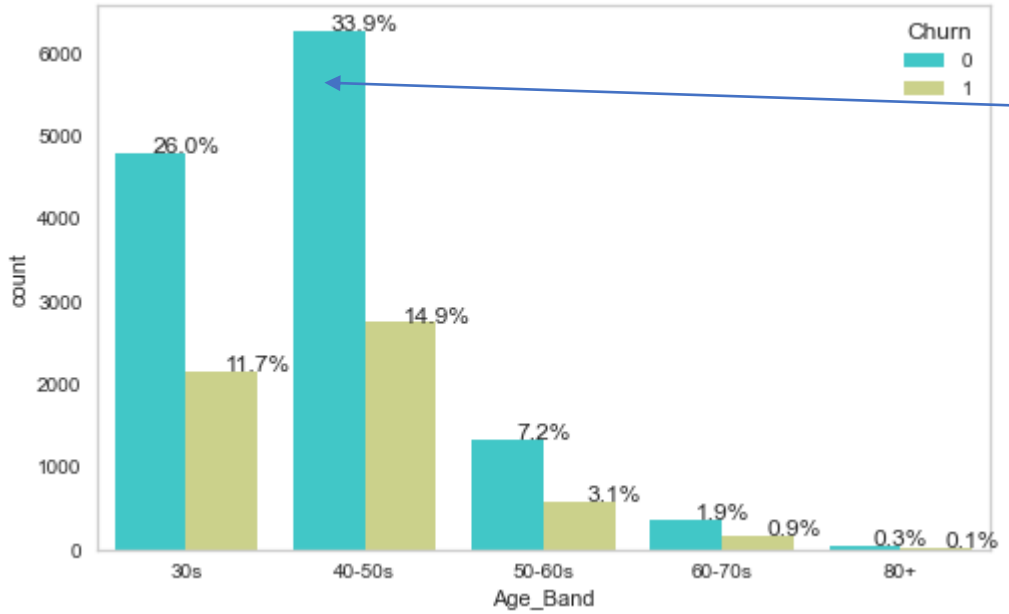


Most customers tend to have bachelor & partial college degrees.

Most customers tend to be Professional or in skilled manual occupation

Exploratory Data Analysis

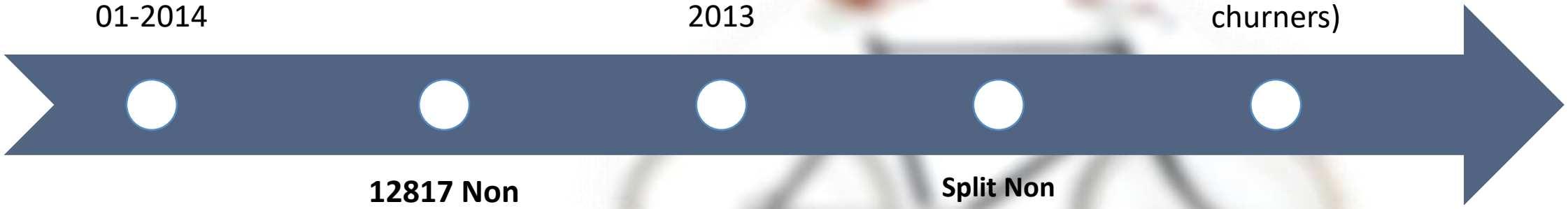
Upon plotting the input variables versus the target variable churn, we are able to visually identify certain key characteristics of the customers



Most customers in the age band of about 40 to 50 years

DATA SELECTION FOR MODEL TRAINING

Out of 18484 customers are 69%(12817) non churners which means they did not purchased products since last 8 months.
This impact profit in over all business.



18484 Total Customer with last order on 28-01-2014

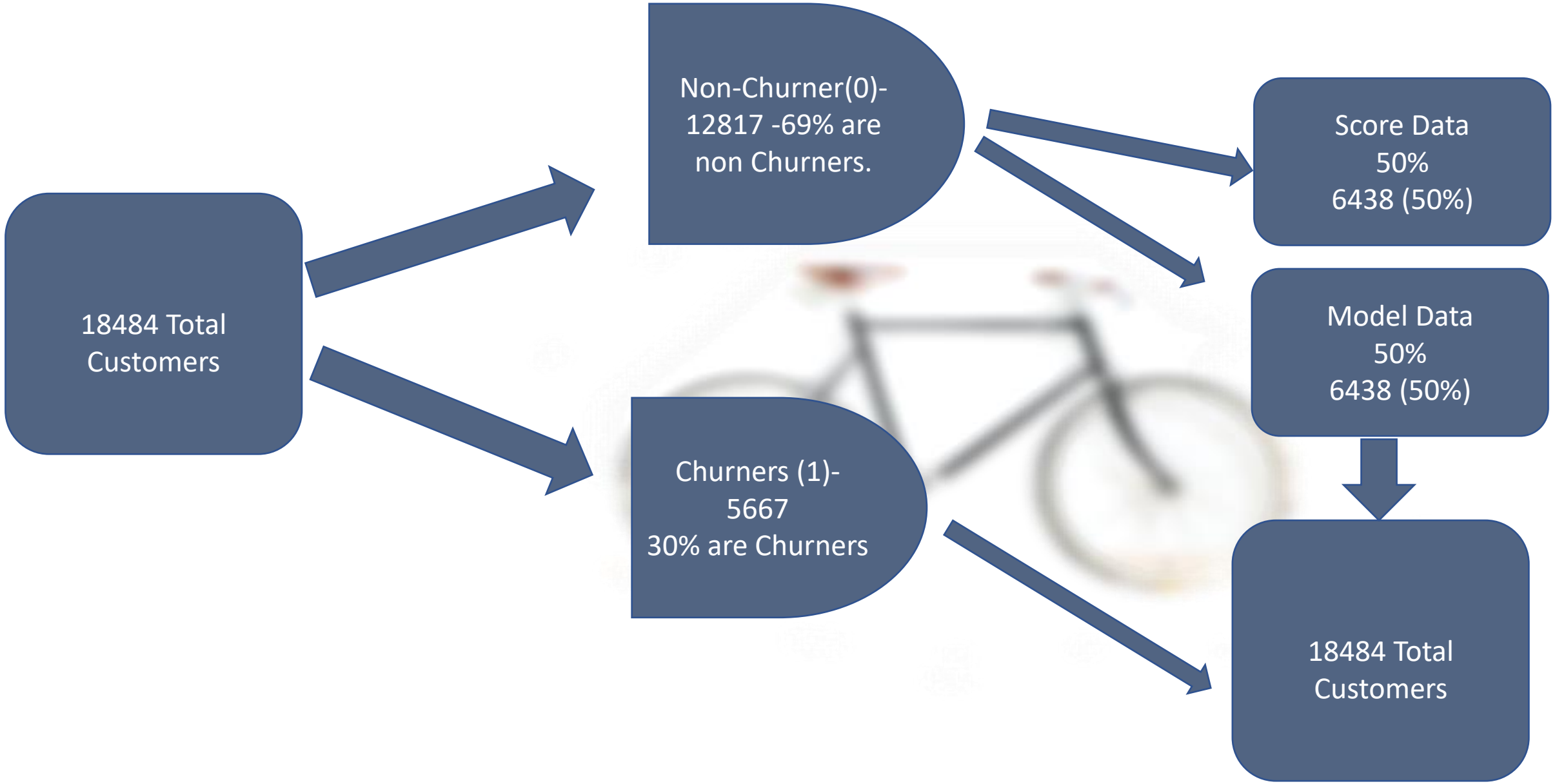
12817 Non Churners. Last purchase after 31-05-2013

5667 Churners Last purchase before 31-05-2013

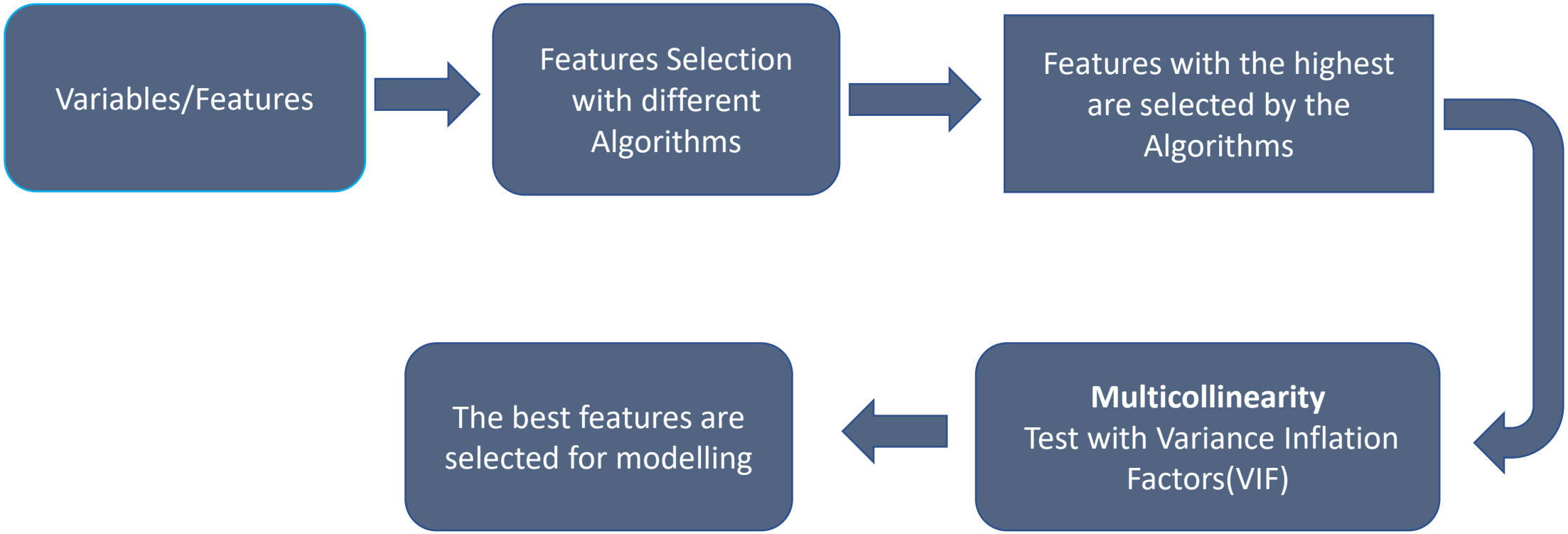
Split Non Churners in 2 parts; 6408 model data & 6408 score data

Train the Model with 12075 customers (6048 model data +5667 churners)

DATA PARTITIONING



FEATURE SELECTION METHODLOGY
VOTE BASED APPROACH



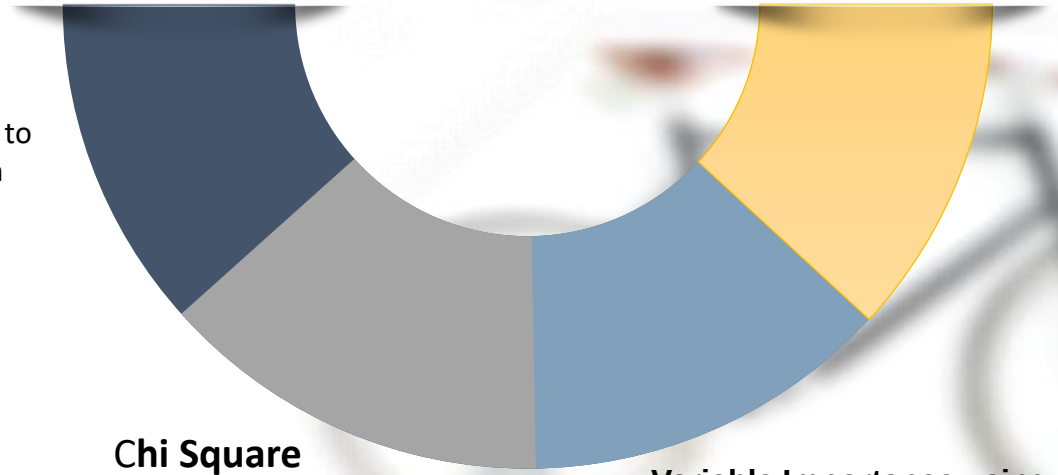
Top reasons to apply feature selection on vote based approach are:

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces overfitting

Variable Selection – Vote Based Approach

Weight of evidence and Information Value

Variable Importance using Random Forest



L1 feature selection

We are going to use regularisation to add penalty in every parameters in the machine learning model, the reason for this is to eliminate overfitting in the model.

Recursive Feature Elimination

eliminates and features from a model by fitting the model multiple times and at each step, removing the weakest features, determined by either the coefficient_or feature_importances_ attribute of the fitted model. The visualization plots the score relative to each subset and shows trends in feature elimination.

Chi Square

We will use chi square to determine which of the features are very relevant to the outcome that will be predicted and remove those that are unimportant for this model

Variable Importance using Extratrees Classifier

Extratrees classifier perform similarly to RandomForest, the only major difference is the level of variance, of course extratrees show the lowest variance.

SELECTED FEATURES FOR MODELING

RFM_Score_11

RFM_Score_8

Frequency

Tenure_Months

RFM_Score_12

Revenue_min

RFM_Segment_442

As per previous slide, with the described method the 9 features were selected for modelling.

RFM_Score_5

RFM_Segment_421

RFM_Segment_443

Model selection

	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
0	Gradient Boosting	0.83	0.87	0.74	0.80	0.76
1	Neural Network	0.81	0.87	0.69	0.77	0.72
2	Logistic Regression	0.81	0.84	0.75	0.79	0.76
3	Decision Tree	0.82	0.81	0.82	0.81	0.82
4	Extra Trees	0.82	0.80	0.82	0.81	0.82
6	Random Forest	0.82	0.79	0.83	0.81	0.82
5	Naive Byes	0.77	0.74	0.77	0.75	0.76

- 80% of the predictive model work is done so far. To complete the rest 20%, we split our dataset into train/test and try a variety of algorithms on the data and pick the best one.
- The model with the best accuracy will be selected, Gradient Boosting has 83 %accuracy which makes it the model for the project

Model Performance & Evaluation

Classification Report

The average precision is 0.83
The average recall is 0.86
The f1-score is 0.85

	precision	recall	f1-score	support
0	0.83	0.86	0.85	1930
1	0.84	0.79	0.82	1693
accuracy			0.83	3623
macro avg	0.83	0.83	0.83	3623
weighted avg	0.83	0.83	0.83	3623

Confusion Matrix

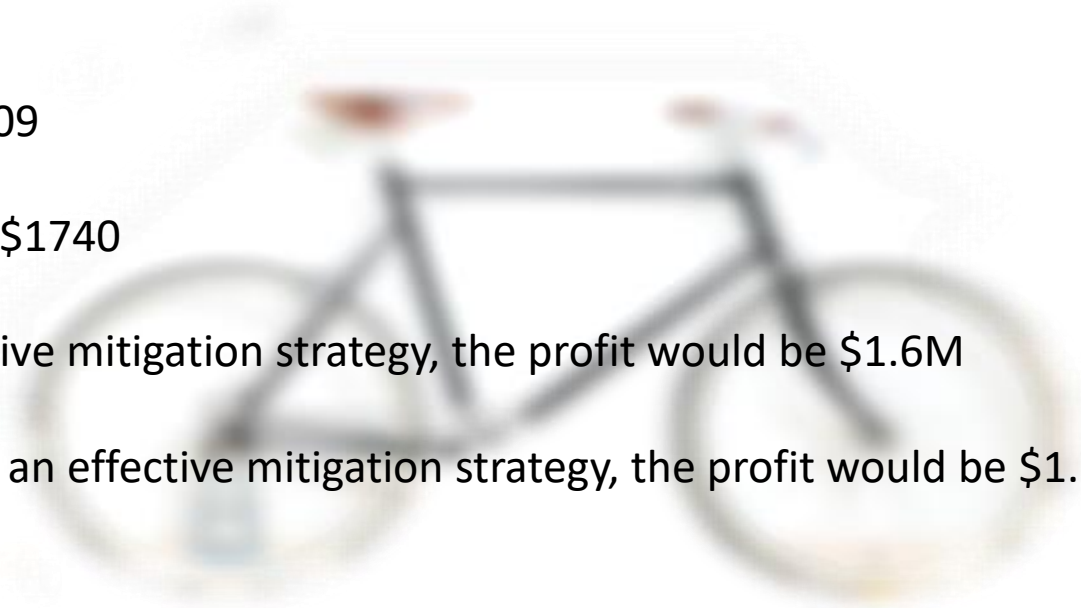
Classified correctly as Non - Churners – 1669
Classified correctly as Churners – 1344
Classified incorrectly as Non - Churners – 261
Classified incorrectly as Churners - 349

0	1669	261
1	349	1344

- The performance of the classifier(GB) is evaluated using confusion matrix by examining the number of observations that are correctly and incorrectly classified. The final model confusion matrix is shown below:

MODEL SCORING & COMMERCIAL IMPACT

- Total number of live (Current) customer – 6409
- Using the selected champion model i,.e Gradient Boosting, the customers likely to churn are 841 and not likely to churn are 5548.
- Total Number of Customers Scored are 6409
- Average revenue per customer per year- $1740
- If we targeted all customers with an effective mitigation strategy, the profit would be $1.6M
- If we target only predicated churners with an effective mitigation strategy, the profit would be $1.7M.This is an improvement of 5.88%.
- Commercial value if successful in preventing churning -$ 2.4M



Recommendation Engine

- Subject to improved result integrate recommendation on online retail platform to flag top three recommendation to each customers

- Revisiting of recommendation engine and carry out test repeatedly make sure that collaborative model is still maintain its relevance over popularity model

- Customized email/SMS for failed or incomplete transactions for all online

- Tightly integrate with marketing team aimed to lure customers with high propensity to churn prospects

Churn Predictive Model

- **Identify customers that are likely to churn** on a weekly basis. Any customer that has not purchased items over the last 6 months is to be classified as a churner or non-churner
- Upon identifying a potential churner, **email/post a promotional voucher or discount coupon**
- Target all potential churners with an effective mitigation strategy of $207K to be prevent them from churning and the average revenue would be at the tune of $223K
- The return on investment would be 448.51% when all the predicted churners are targeted with an effective mitigation strategy
- Products should be recommended to loyal customers often perhaps with some discount to boost cross selling and increase sales and maximize profit.

Thank You

Thank you my Family for continuous support and encouragement.

Thank you Pairview LTD for being patient and continuous encouragement.