Pairview practical Work Experience Project

Data science

21st August 2021

Submitted by Oluseye Oyeniran

Adventure Hardware Group

Reporting Dashboard . **Recommendation Engine** . **Predictive Model**

Executive Summary

- The 3 tools built in this project will empower the management of AHG to make data driven decisions, and in-turn increases the revenue generation in the company
- 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 improve the revenue generated by the company up to £1,129,897,000 every year, if the customers who are likely to churn are prevented from churning as suggested by our model is implemented

1

Reporting Dashboard

- **Purpose of the dashboard**
- **Dashboard features – Filters & KPI's**
- **Dashboard screenshot**

2

Recommendation Engine

- **Need for a recommendation engine**
- **Data transformation, model training, evaluation and selection**
- **Model Output**

3

Predictive Model

- **Purpose of a predictive model**
- **Data exploration**
- **Model training, evaluation and selection**
- **Commercial impact of churn**
- **Recommendations to prevent churn**

1. Reporting Dashboard

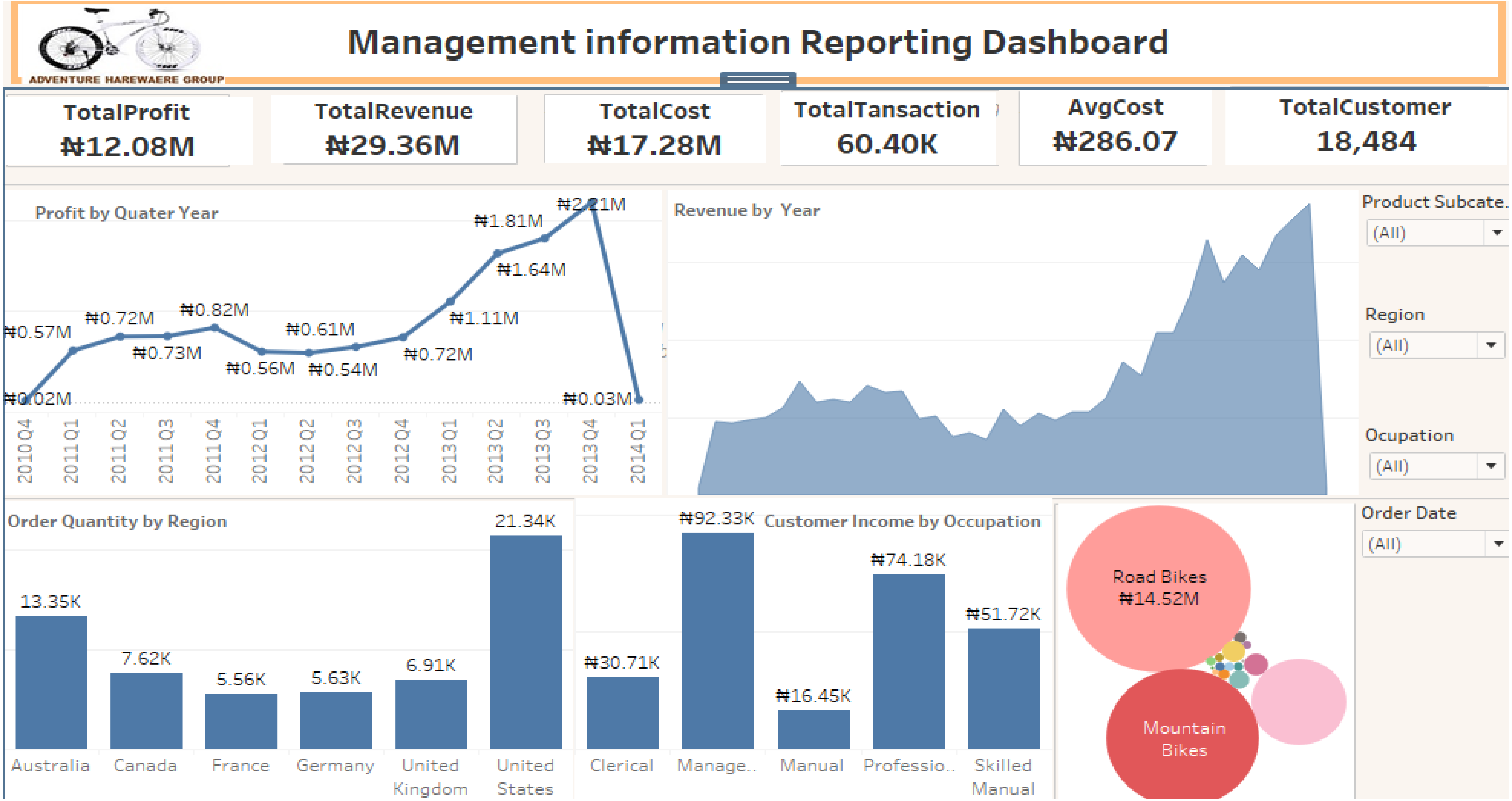
KPI

Overall Performance

- **Total Profit**
- **Total Revenue**
- **Total Cost**

Count

- **Total Transaction**
- **Total customers**
- **Average cost**


ADVENTURE HAREWAERE GROUP
Management information Reporting Dashboard
TotalProfit
₦12.08M
TotalRevenue
₦29.36M
TotalCost
₦17.28M
TotalTansaction
60.40K
AvgCost
₦286.07
TotalCustomer
18,484
Profit by Quater Year
₦0.02M
₦0.57M
₦0.72M
₦0.73M
₦0.82M
₦0.56M
₦0.61M
₦0.54M
₦0.72M
₦1.11M
₦1.64M
₦1.81M
₦2.21M
₦0.03M
2010Q4
2011Q1
2011Q2
2011Q3
2011Q4
2012Q1
2012Q2
2012Q3
2012Q4
2013Q1
2013Q2
2013Q3
2013Q4
2014Q1
Revenue by Year
Product Subcate..
(All)
Region
(All)
Ocupation
(All)
Order Date
(All)
Order Quantity by Region
13.35K
7.62K
5.56K
5.63K
6.91K
21.34K
Australia
Canada
France
Germany
United Kingdom
United States
Customer Income by Occupation
₦30.71K
₦92.33K
₦16.45K
₦74.18K
₦51.72K
Clerical
Manage..
Manual
Professio..
Skilled Manual
Road Bikes
₦14.52M
Mountain Bikes

2. Recommendation Engine

Need for a Recommendation Engine

- AHG has been selling bikes online over the last 4 years. Last year they introduced clothes and accessories as new categories online
- Online sales are account for 1/3rd of the Net Revenue and 11.8 million net profit

The key focus now is to increase each customer basket value and in turn, increase the revenue and profits from the online channel

The tool will be able to search for a recommendation list based on a specified user, such that:

Input: customer ID

Returns: ranked list of items (product IDs), that the user is most likely to want to put in his/her (empty) “basket”

Approach

Building a Collaborative Filtering Recommender System with implicit feedback

Collaborative filtering Explained

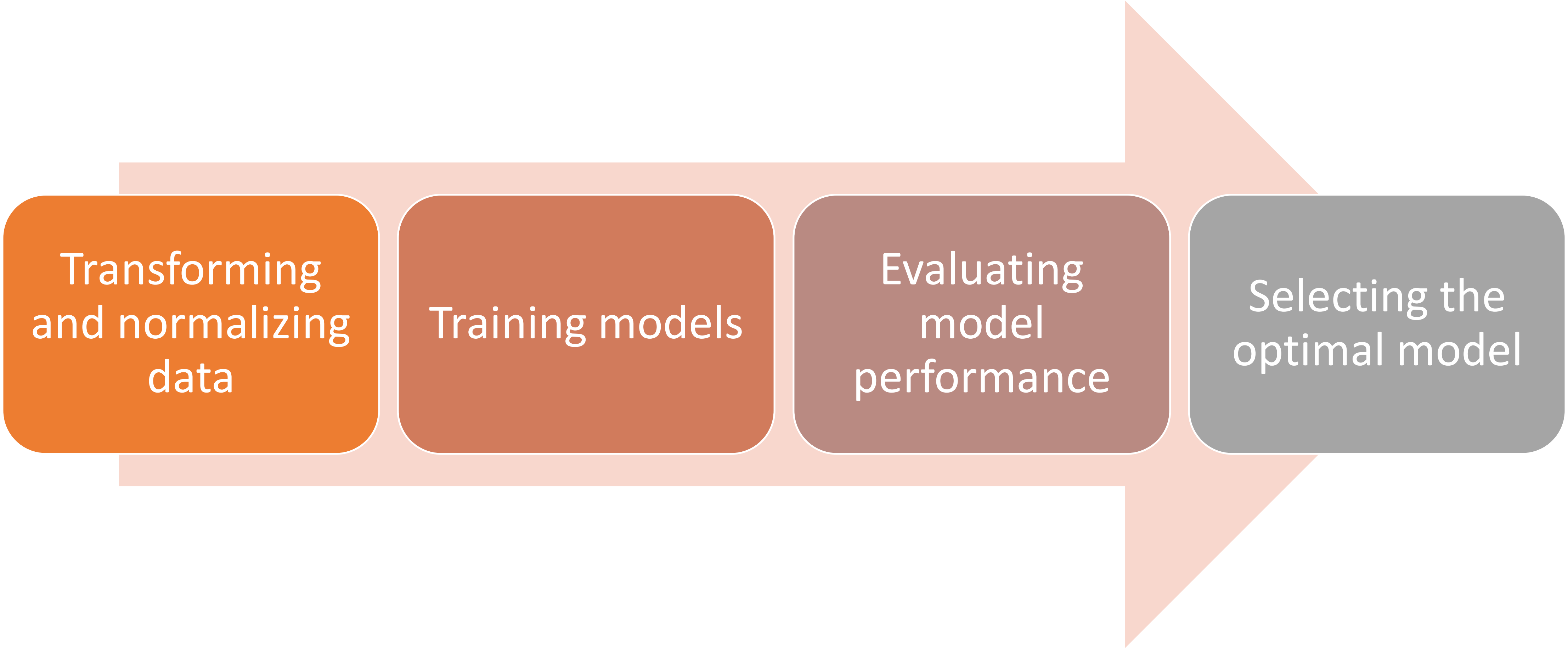
Collaborative Filtering (CF) is a method of making automatic predictions about the interests of a user by learning its preferences (or taste) based on information of his engagements with a set of available items, along with other users' engagements with the same set of items. in other words, CF assumes that, if a person A has the same opinion as person B on some set of issues X={x1,x2,…}, then A is more likely to have B's opinion on a new issue y than to have the opinion of any other person that doesn't agree with A on X. It is a technique that can filter out items that a user might like on the basis of reaction by similar users it works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user

Implicit vs explicit feedback

- Let’s face it, explicit feedback data is hard to collect as they require additional input from the users. The users give explicit feedback only when they choose to do so. As a result, most of the time, people don’t provide ratings at all (I myself totally guilty of this on Konga or Jumia popular e-commerce sites in Nigeria). Therefore, the amount of explicit data collected are extremely scarce.

- On the other hand, implicit data is easy to collect in large quantities without any effort from the users. The goal is to convert user behavior into user preferences which indirectly reflect opinion through observing user behavior. For example, a user that bookmarked many articles by the same author probably likes that author.

Methodology



Data loading, exploration and Transformation

The following field are created

Item decription,Customer ID, **Product ID** and **Order Quantity**

The customer purchase data was transformed as below;

- Group purchase quantities together by Product ID and item ID
- Change any sums that equal zero to one (this can happen if items were returned, but we want to indicate that the user actually purchased the item instead of assuming no interaction between the user and the item ever took place)
- Only include customers with a positive purchase total to eliminate possible errors

Model Training

Transforming and normalizing data → Training models → Evaluating model performance → Selecting the optimal model

Alternating Least Squares Recommender Model Fitting

initialize the Alternating Least Squares (ALS) recommendation model. Fit the model using the sparse_product_customer matrix.

Evaluating Model Performance

Model Evaluation

For evaluating recommendation engines AUC (area under the Receiver Operating Characteristic curve) has been used

AUC (Area under the Receiver Operating Characteristic curve)

- An excellent model has AUC near to the 1 which means it has a good measure of separability. An AUC of 0.77 means the system is recommending items the user in fact had purchased in the test set far more frequently than items the user never ended up purchasing

Commercial Impact of the Recommendation Engine

- To estimate the commercial impact of the recommendation engine, the whole transaction data of the customers was used
- A customer with Id 11000 is chosen for estimate.
- The 4 items recommended to the user from the recommendation model was then quantified in terms of volume, revenue and profit.
- The estimated additional revenue generated per year is **$16,344,928.**

The estimated profit from using the recommendation model per year is **$1,475,840**

Item initially purchased by the customer

	Product_ID	Item_Description
5432	353	Mountain-200 Silver, 38
5442	214	Sport-100 Helmet, Red
5457	541	Touring Tire
5458	530	Touring Tire Tube
5459	573	Touring-1000 Blue, 46
5476	485	Fender Set - Mountain

Items recommended by our model for the customer with ID-11000

	Product_ID	Item_Description
0	358	Mountain-200 Black, 38
1	362	Mountain-200 Black, 46
2	361	Mountain-200 Black, 42
3	478	Mountain Bottle Cage
4	352	Mountain-200 Silver, 38
5	563	Touring-1000 Yellow, 54
6	487	Hydration Pack - 70 oz.
7	562	Touring-1000 Yellow, 50
8	576	Touring-1000 Blue, 60
9	465	Half-Finger Gloves, M

Economic impact of the Recommendation System

If the customers ended up purchasing 4 items out of 10 recommended by our model we then have what is shown below

Item purchase	Total Order Qty	Revenue	Profit($)	Revenue/ orderqty($)	Profit/ Ordrqty($)	
All purpose bike stand	249	39591	24783	159	99	
Mountain bottle cage	2025	20229	12663	10	6	
Mountain-200 black	528	1294866	59047	2452	112	
Sport 100 helmet, black	2085	72954	45669	35	22	
Grand Total for period of 3 years =				2656	239	
Annual Total				886.	80	

This is an estimate for one customer, if all customers purchase an average of four items per year as recommended by the model, the organization could earn **$16,344,928** in revenue and Profit of **$1,475,840** annually

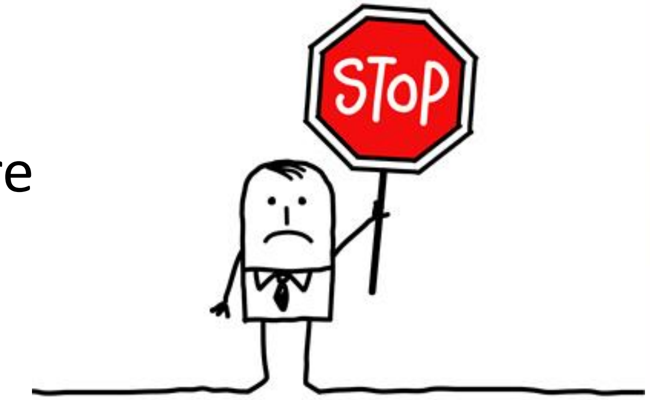
3.Predictive model

Purpose of a Predictive Model

Out of AHG's 18,484 customer database - 5,559 customers have not made a purchase in the last 8 months. These are classified as customers who have churned, and impact the profitability of the business. The management needs quantifiable and timely metrics in order to tackle this.

The Predictive Model will find a pattern in the features of customers who have left in the past, and find similar patterns in existing customers, to send alerts if it finds that a customer is potentially going to leave.

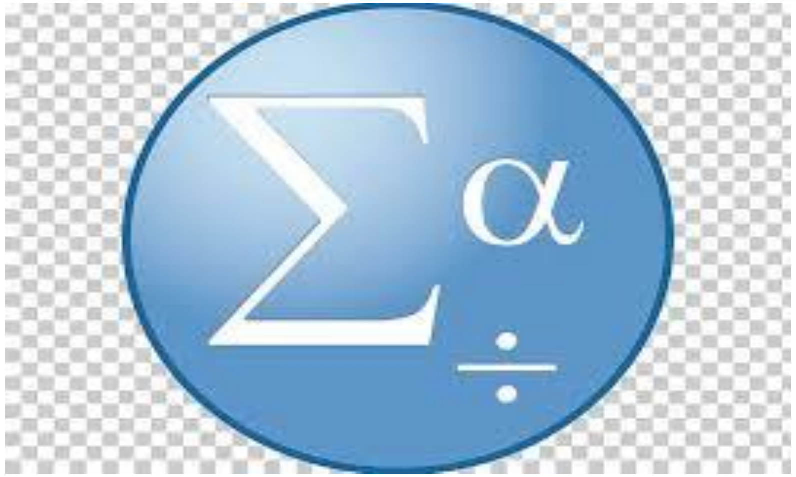
This enables the management to take necessary retention measure



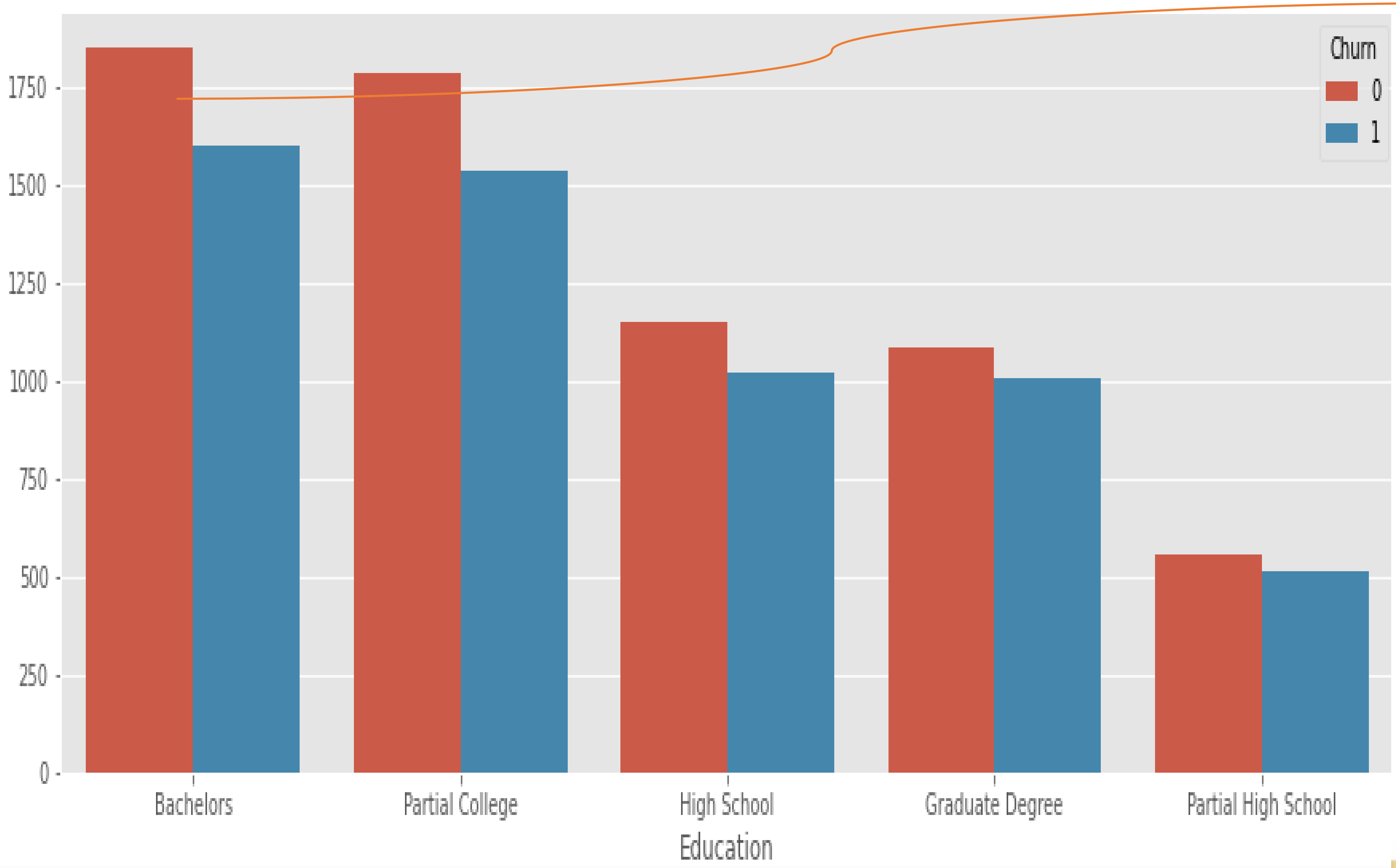
Exploratory Data Analysis

- The target variable is churn, where **1 represent churner,** customer whose maximum order date(last transaction date) is greater or equal to 8 month is assumed to have ceased doing business with AHG (Churner) and **0 is customer who did not (non-churner). 139 features were derived from our data, after performing feature engineering, eight features were selected as the final features to train the model.**

The features are: RFM_cluster_3, RFM_status_Gold,RFM_score_5,RFM_score_8, RFM_cluster_0,RFM_segment_442,Revenue_min

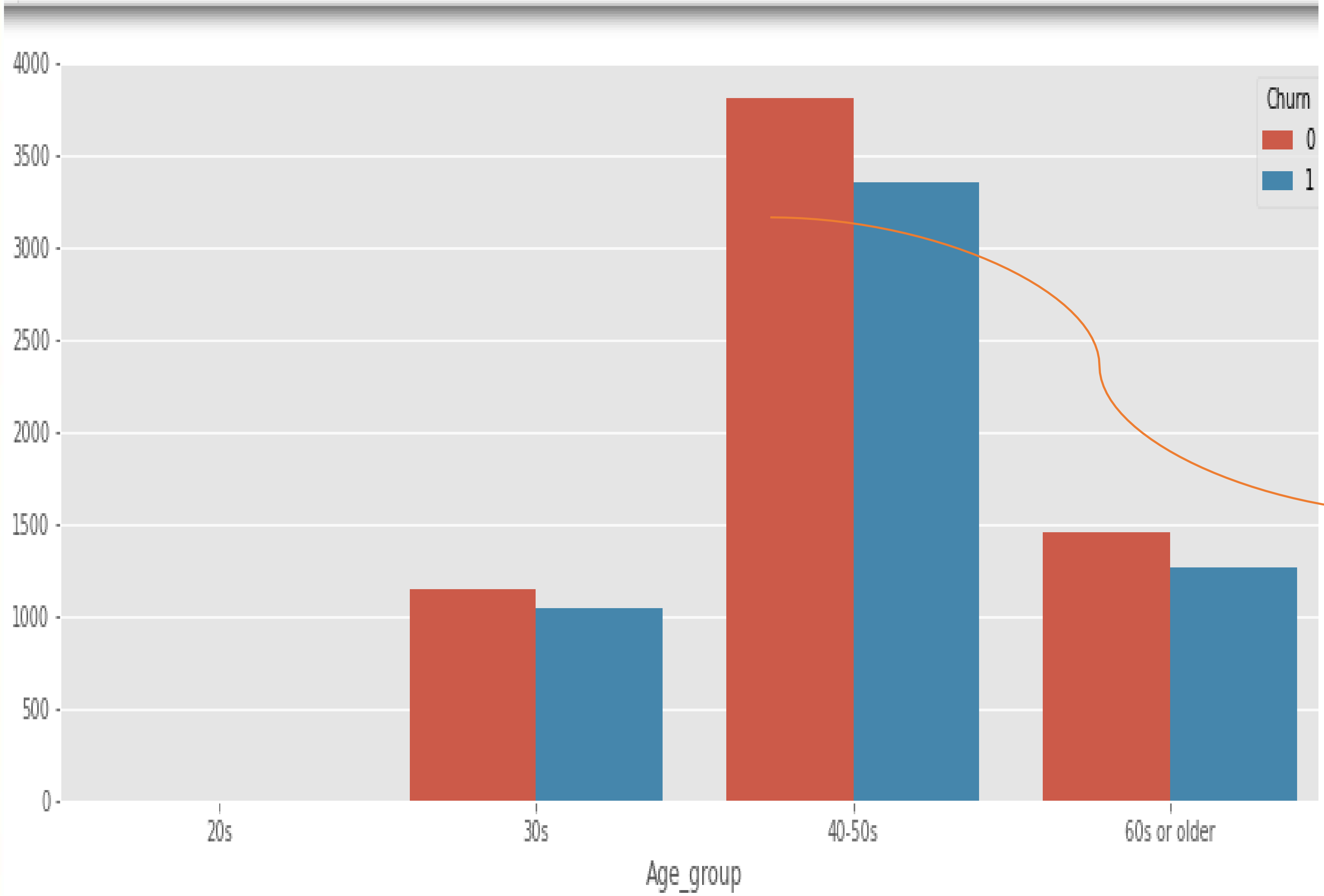


Exploratory Data Analysis: Education?



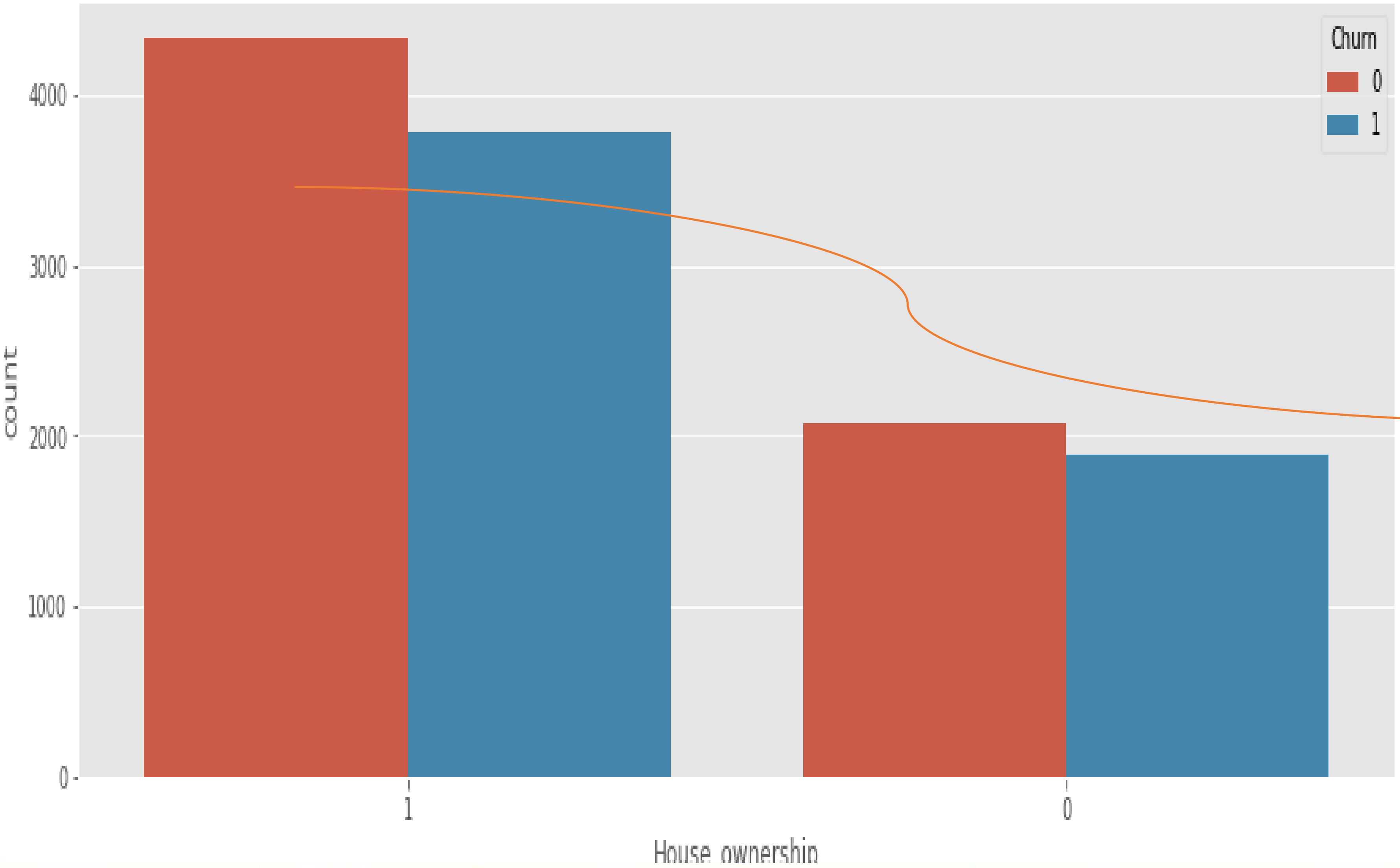
Most of the customers who are loyal (non- churner) have been educated up to a Bachelor degree or partial college

Exploratory Data Analysis: Age group?



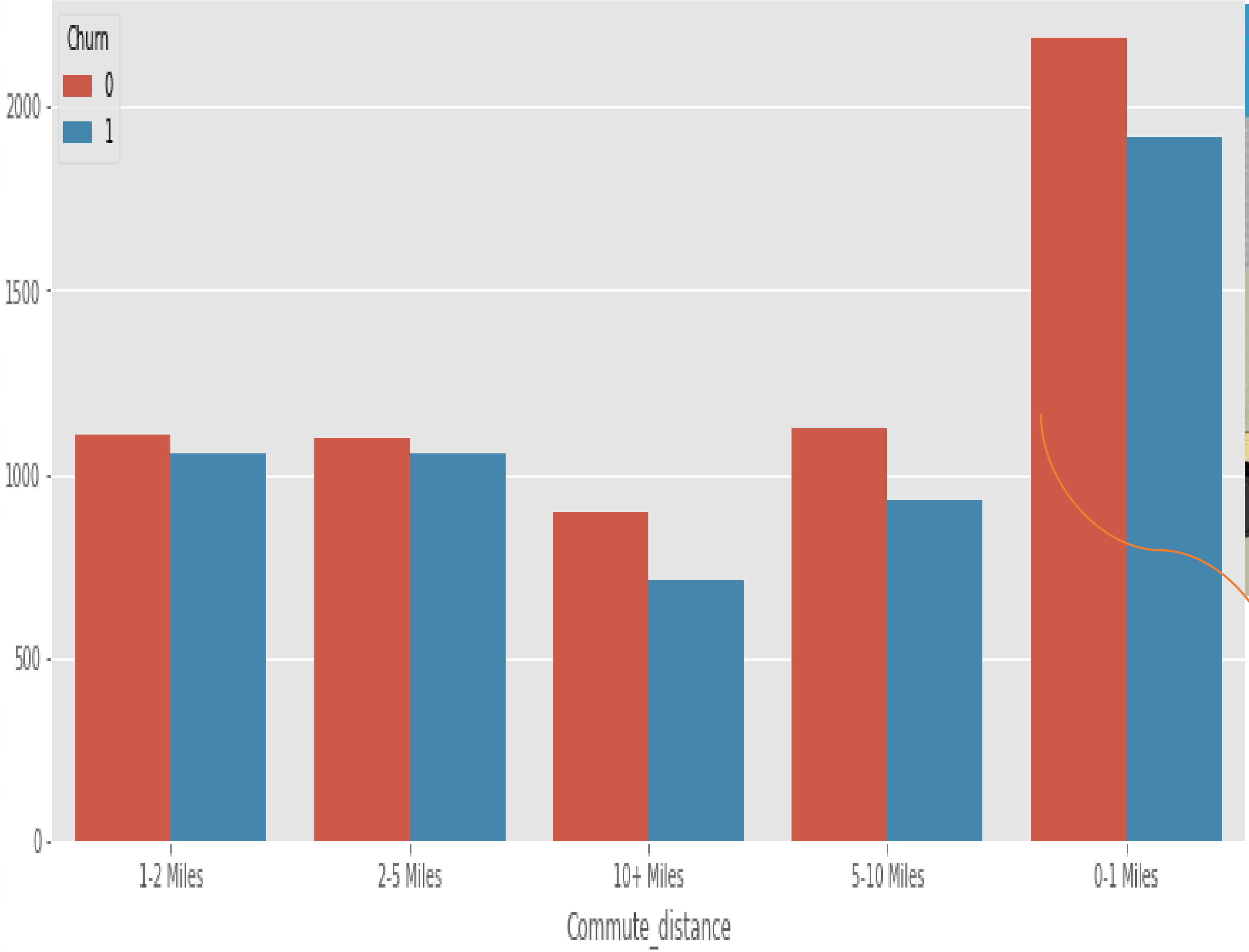
Most of the active customers are within the age bracket of 40 and 50 years, however the loyal customers (non- churners) within that ages are more than the churner

Exploratory Data Analysis: House ownership ?



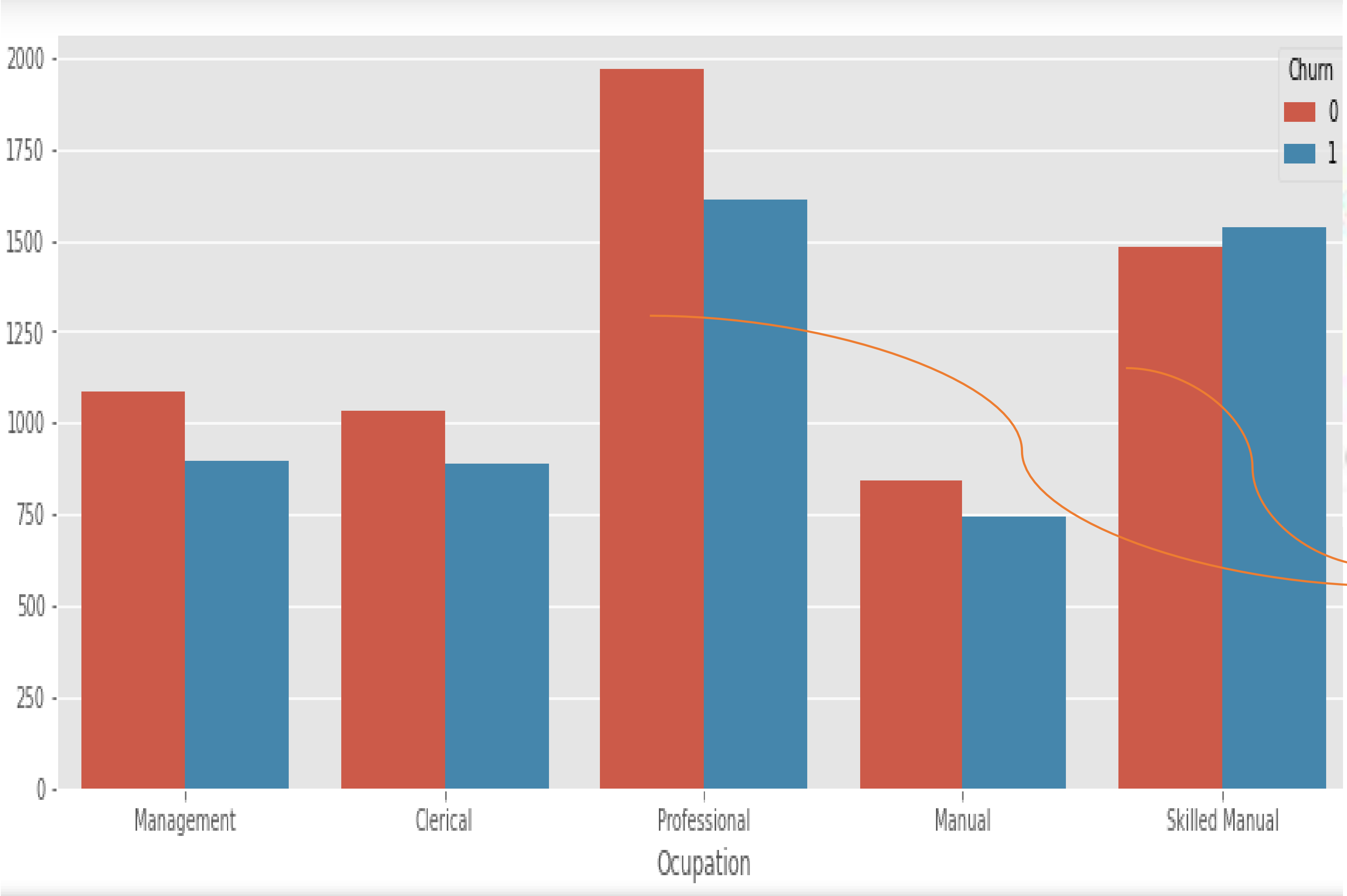
Greater numbers of the customers own house however non churners own more houses than the churners

Exploratory Data Analysis: commute distance?



Greater numbers of the customers commute a distance between 0 and 1 miles however non churners move more distances than the churners

Exploratory Data Analysis: occupation?



Larger proportion of the customers tend to be professionals or in a skilled manual occupation however there are more loyal customers(non-churners) who are in professional occupation than churners in the same occupation

Data selection to train the model

18,484 total customers.

Last order date 28-01-2014

Non Churners – 12,925

Made a purchase after 31-05-2013

Churners – 5,559

Did not make any purchase since 31-05-2013

Non-churners split into 2 parts

6,463 (to train model)

6,462 (to score model)

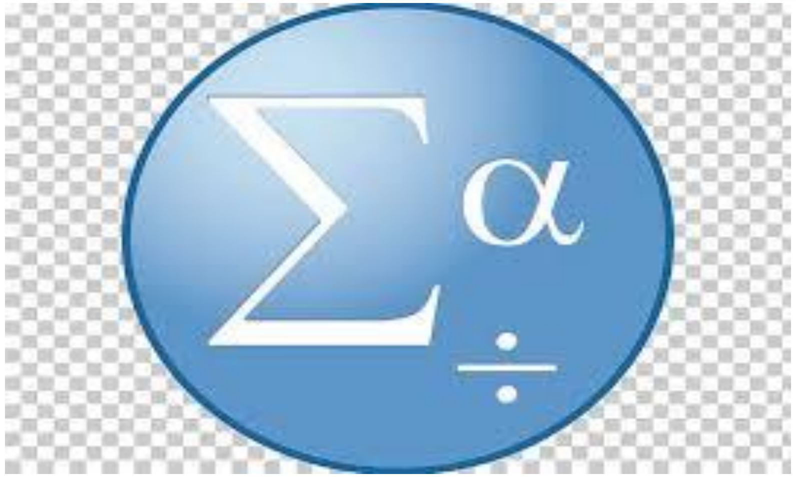
Model trained with

churners 5,559 + non-churners 6,463

= 12,022 customers

Feature selection to train the model

- **139 features were derived from our data, after performing feature engineering, eight features were selected as the final features to train the model. The features are:** RFM_cluster_3, RFM_status_Gold,RFM_score_5,RFM_score_8, RFM_cluster_0,RFM_segment_442,Revenue_min



Models Trained for Predictive Analytics

Logistic Regression

Gradient Boosting

Support vector machine

Naïve Bayes

Logistic Regression

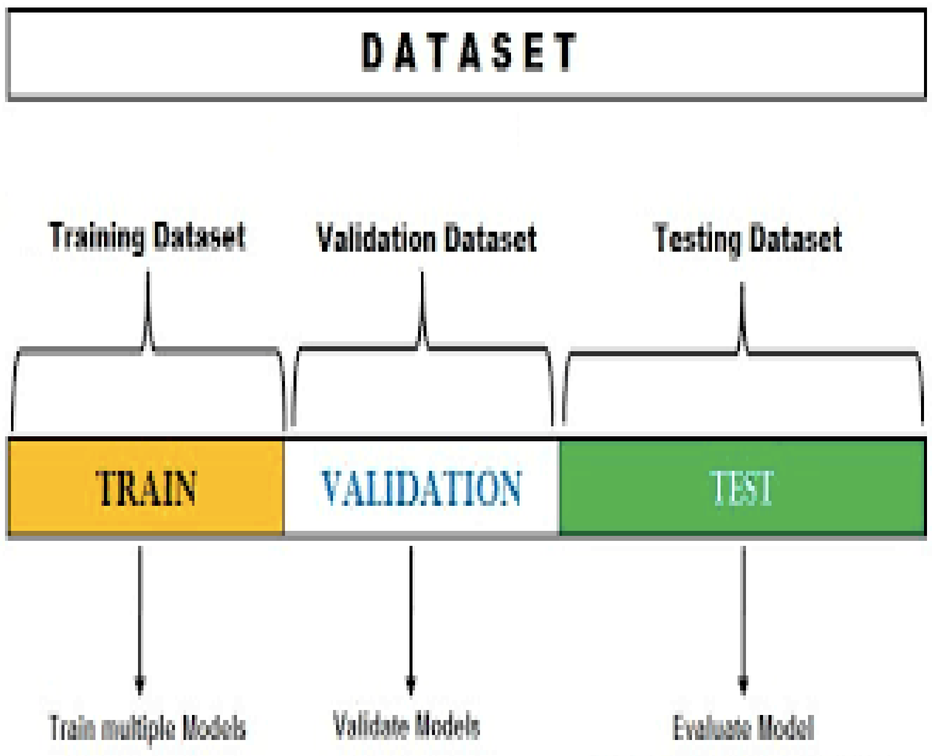
K Nearest Neighbour

Decision Tree Classifier

Neural Network

Linear discriminant Analysis

Random Forest Classifier



The champion model?: Gradient Boosting

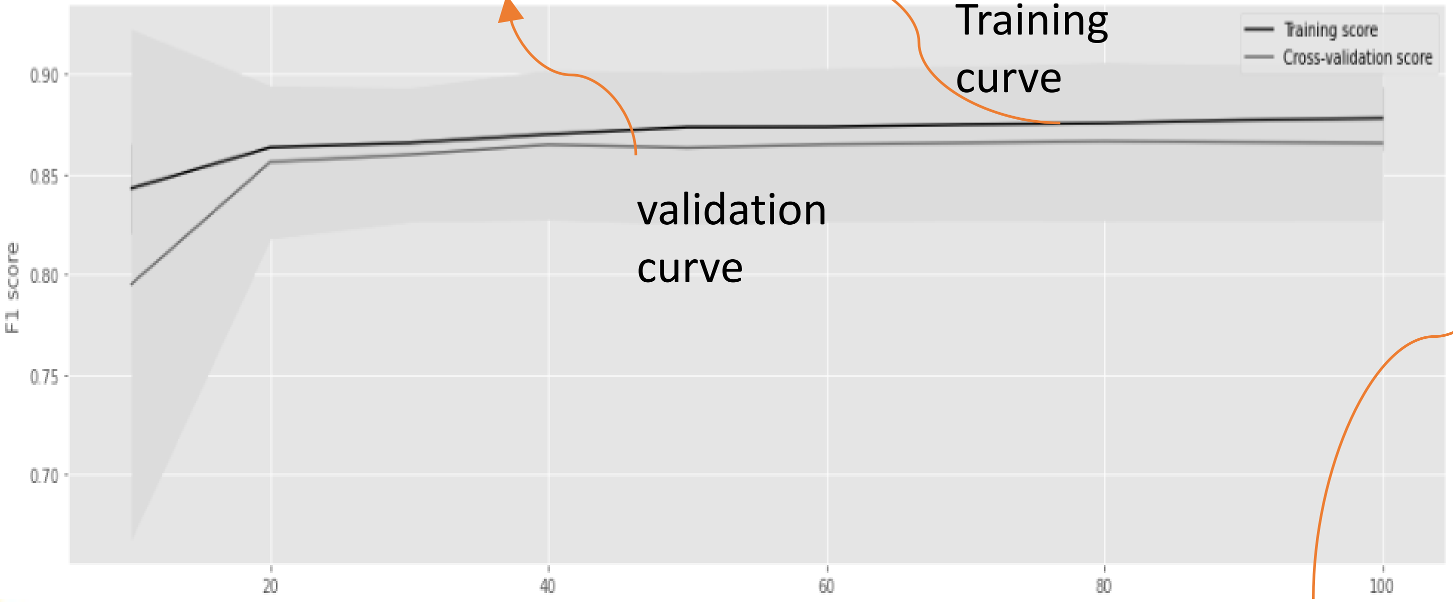
index	test_f1_score
Gradient_boosting	87.64
SVM	87.22
Decision_tree	86.44
Extra_trees	86.40
Random_forest	86.11
KNN	84.48
LDA	83.38
Logistic_regression	83.29
Neural_network	82.80
Naive_bayes	81.22

Gradient boosting is selected as the final winning model because it has the highest f1 score out of the 10 selected models. Because it was able to make accurate prediction on the test set, the data set it has never been exposed to.

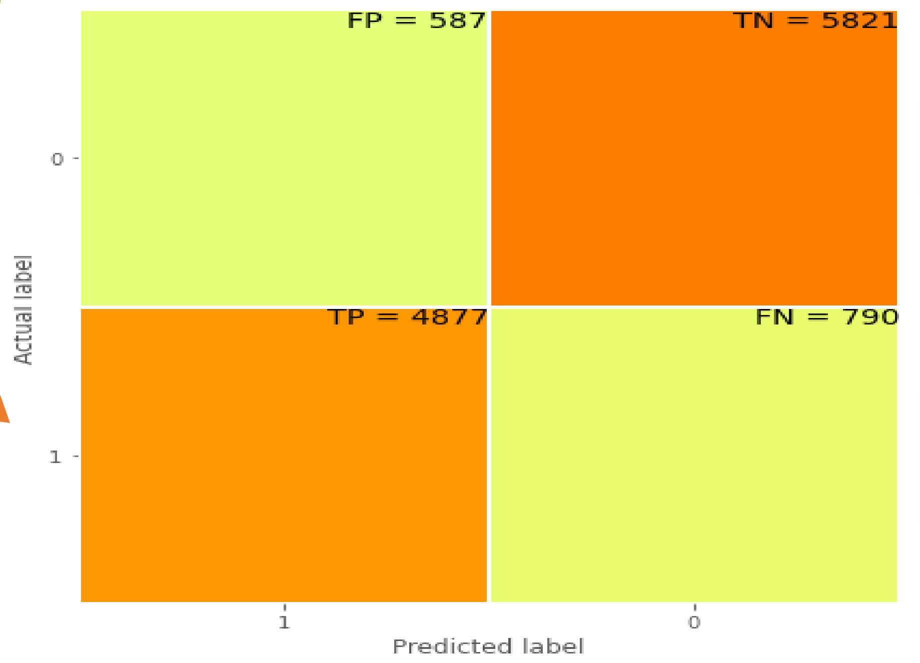
Hyper parameter turning of the gradient boosting Model

Ideally, we would want both the validation curve and the training curve to look as similar as possible..

Our aim is that we would want the value of the parameter where the training and validation curves are closest to each other.



	precision	recall	f1-score	support
0	0.88	0.91	0.89	6408
1	0.89	0.86	0.88	5667
accuracy			0.89	12075
macro avg	0.89	0.88	0.89	12075
weighted avg	0.89	0.89	0.89	12075



We are able to identify an optimized model with a higher accuracy and better performance on the confusion matrix.

The final model has been trained using this. Below are the scores of the final model.

Commercial Impact of Churn

- Customers scored – 6,464
- Customers likely to churn – 572
- Customers likely to stay – 5,837
- Average revenue per customer per year- £1,975

The estimate commercial value if our strategy is successful in preventing churning is $1,129,897

Thank You for your time

Oluseye Oyeniran