Using Machine Learning to to accurately classify customers, establishing a baseline for recommendation engines and personalised marketing strategies

PG CERT PROJECT PROPOSAL 2024/25

Describes the problem area, the aims and objectives of the project, the methodology to solve the problem and the plan towards the proposed solution.

Contents

1. Project Kernel

INSTRUCTIONS FOR STUDENTS **(1)** complete boxes A, B **(2) Upload the completed form by deadline specified on Moodle**	**INSTRUCTIONS FOR STAFF** **(1)** complete boxes C and D

A **Proposal from:** Rodney Sibanda

Email: rsiban01@student.bbk.ac.uk

Reg No: 14016555

B

Suggested working title: Customer experience in retail - How I use data analysis to improve customer experience

General topic area (e.g. Data Analytics, Databases, Machine Learning, Software Development):

Data Science; Software Development

Brief description of problem to be tackled and up to three key references:

Analysis customers experience data from different sources, identify patterns & anomalies that impact on customer experience and then provide actionable insight/ recommendations for eCommerce platform product owners.

Online shoppers perceived impact of AI in customer service in 2024

ACSI - U.S. customer satisfaction with online retail as of 2024

Opinions on the online shopping customer experience 2022

Level of project difficulty (from 1 for "appropriate to the degree" to 5 "for very challenging")**:**
This is 3, integrating/ harmonising the datasets will be the hardest part. I should be able to use Python/ SQL to import the data into a staging environment where I can profile, cleanse and analyse the data using tools I've worked with on the course.
If I can manage that, I can talk to my supervisor about stretch targets.

Resources required: (list dataset, software and hardware)
Visual Studio and open-source modules/ libraries. Other datasets may be considered after discussion with the supervisor.

Are these currently available within the School? Yes, The projects will use open-source tools. I may need to get subscriptions to acquire the datasets

Is there an outside company involved? Yes, my own limited company, Agile Chameleon.

C

This project is approved for the PG Cert Applied Data Science **YES/NO**

D (project tutor) **Status:** **OK** (accepted) **RP** (revise) **RA** (revision accepted)

Supervisor:

2. Introduction

This section will describe the scope & objectives of the project & provide background information explaining its real-world relevance.

2.1 Overview of the project scope and objectives

2.1.1 In scope

The project is to identify the key features/ information a retailer requires to identify a demographic/ class label for a new customer/ prospect.
Classifying that new customer/ prospect will enable retailers to personalise their marketing strategies & improve recommendations.

2.1.2 Out of scope

This project will not cover: -

- Recommendation Engines
- Dynamic Pricing
- Demand forecasting
- Chatbots – customer service
- Visual & Voice Search

These AI Tools enhance customer experience & will be examined in future projects.

2.2 Importance of accurately classifying customers/prospects

The challenge for most B2C businesses is how to engage with customers in a way that increases customer sentiment. A lot of marketing campaigns are flagged by customers as junk mail; recommendations are ignored.
The click-thru-rate for a retail email campaign is between 0.1% to 0.5% of recipients - about 2% of click-thru visitors get converted to a sale (or 0.002% to 0.01% of email recipients ultimately convert to a sale) [1]
In a competitive environment the B2C businesses that can successfully engage with their customers wins.

AI in Retail: How classification drives personalised shopping

Accurate classification of customers enables retailer to predict individual preferences and suggest products aligned to those preferences. It is a core part of any successful recommendation engine.

Unsupervised ML models can cluster customers by analysing a vast amount of customer data — from browsing habits to purchase history. Those clusters can be converted to classes for supervised models and used to predict individual preferences with remarkable accuracy.

Platforms like Spotify and Netflix use personalisation to boost customer engagement, enhance satisfaction, and increase sales.

[1] Updated 2023: Average Conversion Rate by Industry and Marketing Source - Ruler Analytics - Average lead conversion rate for email (2021)

Continuous optimisation/ fine-tuning of B2C classification ML models along with the accumulation of relevant customer data will improve the accuracy of the classification ML model predictions and the effectiveness of the recommendation engines and personalised marketing strategies that use classification ML model outputs[2].

[2] How Spotify Delivers a Unique Customer Experience (CX) with Personalized Music Recommendations
How Netflix, Spotify & TikTok Use Personalized Recommendations
https://hbr.org/2024/11/personalization-done-right

2.3 Key challenges addressed by project

This project will provide tools for B2C businesses to identify the key information required to classify a new customer/prospect.

3. Project Overview (Heatmap)

Schedule Scope Resources

Tasks are marked as follows: on schedule (GREEN), at risk (ORANGE), and behind schedule (RED). Currently, everything is on schedule.

May 5	May 19	Jun 2	Jun 16	Jun 30	July 14	July 28
Prepare Project Plan	1st pass Data Acquisition	Supervisor reviews project proposal	Submit Project Proposal	Fine tune data acquisition/ data cleansing	Fine tune data exploration & data visualisation	Supervisor reviews project/ dissertation
Prepare GitHub Repository	1st pass Data Cleansing	1st pass Data Visualisation		Prepare 1st draft of final dissertation	Update draft dissertation	
	1st pass Data Exploration					
	Update Project Proposal					

Aug 11	Aug 25					
Update data pipeline & dissertation	Submit project and dissertation					

4. Problem Statement

4.1 Existing gaps in customer classification

Over-recommendation & spamming results in poor customer sentiment which is directly linked to a customer's lifetime value and sales conversions from customer engagements. Retailers always need to grow their market share and break into new markets.
For example, a B2C business make most of their sales to Gen X & Baby Boomer customers – a constantly shrinking market. Their long-term survival is dependent on increasing and retaining a larger number of Millennial & Gen Z customers.
A key to penetrating these markets is accurately identifying Millennial & Gen Z customers/prospects and using personalised marketing strategies to engage and retain them.
Personalised marketing strategies are dependent on accurate classification of customers/prospects. The accurate classification of customers/prospects is a problem for many B2C domains – from retail to entertainment, healthcare & financial services.

4.2 Need for improved data-driven insights into consumer demographics

Accurate classification will make it easier to predict what a customer may be interested in.

Prediction is largely dependent on classification and accurate classification is dependent on a large enough sample of the right input features for the target customers/prospects. [3]

Demographics is a type of classification ideal for supervised ML models.
Analysing customer data, for example, past purchases, browsing behavior, and preferences is ideal for unsupervised ML models. Unsupervised ML models can identify clusters and clusters can be converted to classes for supervised models.[4]

AI-native approaches offer a more dynamic and personalised way to leverage vast amounts of data to enhance every aspect of the customer journey. [5]

4.3 Potential impact on retail and e-commerce strategies

Collecting the right input features for a large enough sample of existing customers/ prospects will enable data scientists to train a supervised classification model to identify the correct labels for new customers/ prospects.
Accurately classifying customers/prospects will allow retailers to optimise personalised marketing strategies & recommendation engines in any domain.

[3] 6 Ways AI Can Improve the Customer Experience

[4] 5 ways of enhancing customer experience in retail with AI - BOI (Board of Innovation)

[5] 13 ways AI will improve the customer experience in 2025

5. Data Acquisition

5.1 Source datasets from Kaggle

- Customer Satisfaction Response to AI – ideal for demographic classification
- Shopping Trends – Ideal for classification based on past purchases, browsing behavior, and preferences
- Customer Demographics and Spending – ideal for demographic classification
- Retail Sales Dataset - Ideal for classification based on past purchases, browsing behavior, and preferences
- ***Looking for other datasets***

5.2 Synthetic data

These samples are too small to deliver accurate predictions. Bootstrapping will increase the size of the samples and the accuracy of the model predictions.

5.3 Justification for dataset selection and merging approach

The individual datasets have limited input features. Merging data from different datasets and engineering new features will provide a larger range of input features.

5.3.1 Example of Merging datasets

In order to create the dataset AI_Retail_WIP_Eng I started with the ‘Customer Satisfaction Response to AI’ dataset.
The original dataset contains 656 rows and 23 columns, with all columns having data in string format. Below is a description of each column:

Country: The country of the consumer
Age: Age group of the consumer (e.g., Gen X, Gen Z).
Annual_Salary: The consumer's income level (e.g., Low, Medium High, High).
Gender: Gender of the consumer (e.g., Male, Female).
Education: Educational qualification (e.g., University Graduate, Masters’ Degree).
Living_Region: Type of area the consumer resides in (e.g., Metropolitan, Rural Areas).
Online_Consumer: Indicates whether the consumer shops online (YES/NO).
Online_Service_Preference: Preference for online services (YES/NO).
AI_Endorsement: Indicates trust in AI technology (YES/NO).
AI_Privacy_No_Trust: Indicates concern about AI and privacy (YES/NO).
AI_Enhance_Experience: Whether AI enhances the shopping experience (YES/NO).
AI_Satisfaction: Satisfaction level with AI (YES/NO).
AI_Tools_Used_Chatbots: Whether the consumer uses chatbots (YES/NO).
AI_Tools_Used_Virtual_Assistant: Use of virtual assistants like Alexa (YES/NO).
AI_Tools_Used_Voice&Photo_Search: Use of AI tools for voice/photo search (YES/NO).
Payment_Method_COD: Use of cash on delivery as a payment method (YES/NO).
Payment_Method_Ewallet: Use of digital wallets for payment (YES/NO).
Payment_Method_Credit/Debit: Whether the consumer uses credit/debit cards for payments (YES/NO)
Product_Category_Appliances: Preference for purchasing appliances (YES/NO).
Product_Category_Electronics: Preference for purchasing electronics (YES/NO).
Product_Category_Groceries: Preference for purchasing groceries (YES/NO).
Product_Category_Personal_Care: Preference for purchasing personal care items (YES/NO).

Product_Category_Clothing: Preference for purchasing clothing (YES/NO).

The Customer Satisfaction Response to AI dataset was augmented with data from the Shopping Trends dataset. A matching feature, Mapping, was created in both datasets by concatenating the gender and age attributes. A lookup was then performed to integrate additional data from the Shopping Trends dataset into the corresponding Mapping variables in the Customer Satisfaction Response to AI dataset.
Additional data cleansing was conducted to account for features missing from the Shopping Trends dataset. For instance, Purchase Amount was adjusted to incorporate the Annual Salary feature from the Customer Satisfaction Response to AI dataset, recognising that individuals with higher salaries generally spend more than those with lower salaries.

Item Purchased - The item purchased (Integer)
Category - Category of the item purchased (String)
Purchase Amount - The amount of the purchase (Integer)
Size - Size of the purchased item (String)
Color - Color of the purchased item (String)
Season - Season during which the purchase was made (String)
Review Rating - Rating given by the customer for the purchased item (Float)
Shipping Type - Type of shipping chosen by the customer (String)
Previous Purchases - The total count of transactions concluded by the customer, excluding the ongoing transaction (Integer)
Frequency of Purchases - Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

Methodology

The methodology will focus on how we can optimise/ fine-tune several different supervised ML models by tweaking their parameters to improve the accuracy of their predictions.

It will also look at how we can take a large amount of customer data without classes or labels. Use unsupervised ML models to identify clusters and convert those clusters to classes for supervised ML models.

5.4 Cleaning and encoding categorical data

Encoding will be required, especially for unsupervised data models.
Supervised models will identify and focus on the most important input features in the data model.

6. Data Modelling & Evaluation – Unsupervised ML Models & Fine-tuning Supervised ML Models

6.1 Link to GiTHub with Jupyter files showing experiments

danbasi-afk/PG_Cert_Project_RS_Repository

6.2 K-means algorithm

The K-Means algorithm was utilised to gain deeper insights into the dataset that may not be accessible through supervised machine learning models.

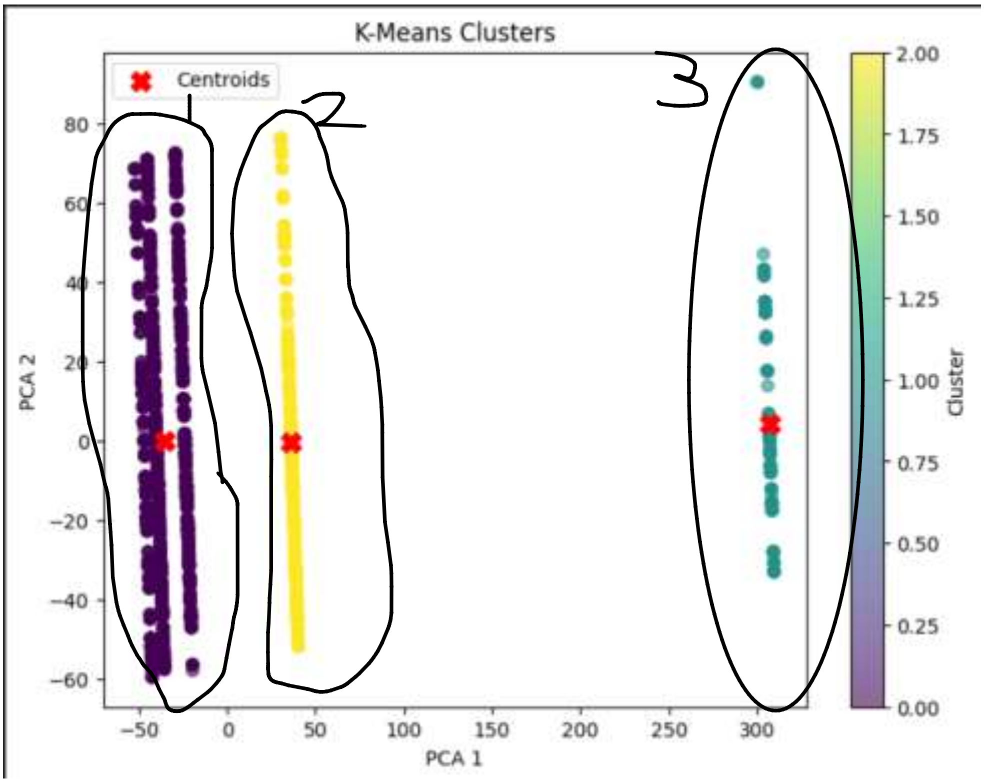
This model used the input features to identify clusters, those clusters will be converted to classes in the Project Report.

6.2.1 *Data Cleansing*

All the features for the K-Means test required encoding before analysis. Once the encoding process was completed, six features out of a total of 38 were chosen for the experiments.

6.2.2 *Exploration & Visualisation*

Based on the silhouette score (0.48), the clustering appeared to be effective. The optimal number of clusters was 3, more than 3 reduced the silhouette score. When visualised, the data formed three distinct groups, each with clear boundaries and well-defined centroids.



To further analyse the clusters, the outputs were exported to excel for examination.

6.2.3 *Variance Analysis*

The variance for each feature was examined to determine its impact on the performance of the model. Features with the highest variance contributed the most to the effectiveness of the clustering process.

6.2.4 *Summary*

This type of unsupervised model would be ideal if attempting to cluster customers based on features like purchase amounts and frequency of purchases. Standardisation techniques, like Standard Scaler, can be employed to normalise features when high variance negatively impacts model performance.

6.2.5 *Next Steps*

Try to use other clustering algorithms like DBSACAN and hierarchical clustering to see if they provide additional insights.

Convert clusters to classes in the Project Report. Train Supervised ML Models to accurately predict the classes for new customers/ prospects based using this data.

6.3 Parameters tuning for the different algorithms

Attempted to improve the performance of the different models by fine-tuning the parameters of each one.

I used cross-validation to evaluate six different ML models to determine which one would most accurately label a new data record, the label was ‘Segment’. This provided the control or baseline results for the fine-tuning experiment.

Then I fine-tuned the parameters for each of these models to see what impact it would have on the accuracy of the results.

The models used are:

- Decision Tree Classifier
- KNeighbors Classifier
- Linear Discrimination Analysis
- XGBoost
- GaussianNB
- Logistic Regression

6.3.1 *Linear Discriminant Analysis*

- The control experiment **result is 12%**
- Increase the number of splits – this is recommended for improving the performance of all the models. Not viable in this instance, there is a class with only 2 members so I can’t have more than 2 splits.

- Standardise the features – prevents features with a large numerical range from overshadowing features with a small numerical range. The **result is 13%**
- LDA Solver – tried both 'lsqr' & 'eigen' – it appears the sample size is large enough that this did not make a material difference. Included the shrinkage parameter to reduce overfitting. The **result is 13%**

6.3.2 *Decision Tree Classifier*

- The control experiment **result is 90%**
- Increase the number of splits – this is recommended for improving the performance of all the models. Not viable in this instance, there is a class with only 2 members so I can't have more than 2 splits.
- RandomForestClassifier – increases accuracy by increasing the weight of the most important features (in this case the top 5 features). Experiments with this did not deliver a material change. The **result is 90%**
- GridSearch to find the best combination of parameters provided for the model – no combination of decision tree parameters made a material difference to the accuracy of the model. The performance of the RandomForestClassifier is measurably better. The **result is 80%**

6.3.3 *XGBoost*

- The control experiment **result is 90%**
- Attempted a Randomized Search instead of Grid Search to find the best combination of parameters provided for the XGBClassifier. There was a marginal improvement with one combination. The **result is 90%**
- Standard Scaler – subtracts the mean from each feature normalising the data to improve model performance. Experiments with this did not deliver a material change. The **result is 90%**

6.3.4 *Logistic Regression*

- The control experiment **result is 12%**
- Regularisation – used a low value to improve generalisation. Experiments with this did not deliver a material change. The **result is 12%**
- Solver – tried different solvers to see which had the most impact on the performance of the model.
  - 'liblinear' – Delivered a measurable improvement to the performance of the model. The **result is 14%**
  - 'newton-cg' – There was some improvement but the 'liblinear' performed better. The **result is 13%**
- Penalty - only the 'l2' worked with the Logistic Regression Model used for this experiment. Experiments with this did not deliver a material change. The **result is 12%**
- Standard Scaler – subtracts the mean from each feature normalising the data to improve model performance. Experiments with this did not deliver a material change. The **result is 13%**

6.3.5 GaussianNB

- The control experiment **‘mean test score’ is 13%**
- Attempted a combination of, ‘var_smoothing-le-1’, and a Standard Scaler. There was negligible improvement in the performance of this model. The **result is 15%**
- Tried the GaussianNB version of the RandomForestClassifier – ‘feature_selection’ – selected the most important features (in this case the top 5 features). The **result was 14%.**

6.3.6 KNeighbor Classifier

- The control experiment **‘mean test score’ is 68%** with n_neighbor=3
- GridSearch to find the optimal n_neighbor value for the model between 1 and 20. 1 was identified as optimal value it had a marked impact on the performance of the model – **from 68% to 90%.**
- Used the Manhattan metric instead of the Euclidean, the default. It made no material difference even after I added feature scaling. **Result is 69%**
- I also used weight=’distance’ instead of giving equal weight to all the nearest neighbors. The **result to 90%** - virtually the same as changing n_neighbor=1.

6.3.7 Summary

The most noticeable improvement was changing the n_neighbor value from 3 to 1. This resulted in a 20% improvement in accuracy, as did weights=’distance’.

The other noticeable improvement was with solvers on the Logistic Regression Model. Accuracy improved by 2% when the ‘liblinear’ value was used with the solver parameter.

The XGBClassifier was highly accurate before attempting to fine-tune the parameters. Randomised Search identified a combination of parameter values that improved performance of this model by 0.1%.

Fine-tuning parameters can improve performance; how much impact is dependent on the model and the dataset.

6.4 *Next Steps*

Apply what I have learnt about parameter fine-tuning to optimize the performance of ML Models used in future experiments for the final project.