Repeat Buyers Prediction

CECS 550 : Pattern Recognition
Spring 2023

Group :- 7

Anthony Martinez	012500759
Diksha Patil	030849765
Pratik Jadhav	030880471
Pavan More	030873555
Sudarshan Powar	030845787

Instructor: Prof Mahshid Fardadi
Teaching Assistant: Rahul Deo Vishwakarma

CALIFORNIA STATE UNIVERSITY
LONG BEACH

Agenda

- Background and Goals
- Data Interpretation and Visualization
- Feature Engineering and ranking
- Prediction models
- Performance
- Recommendations
- Conclusion

Background and Goals

1. A shop runs big promotions on “Double 11” - the biggest online shopping event, in order to attract a large number of new buyers.

2. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may barely a have long-lasting impact on sales.

3. To reduce the promotion cost and enhance the return on investment (ROI), they want to identify who can be converted into repeated buyers.

Goals

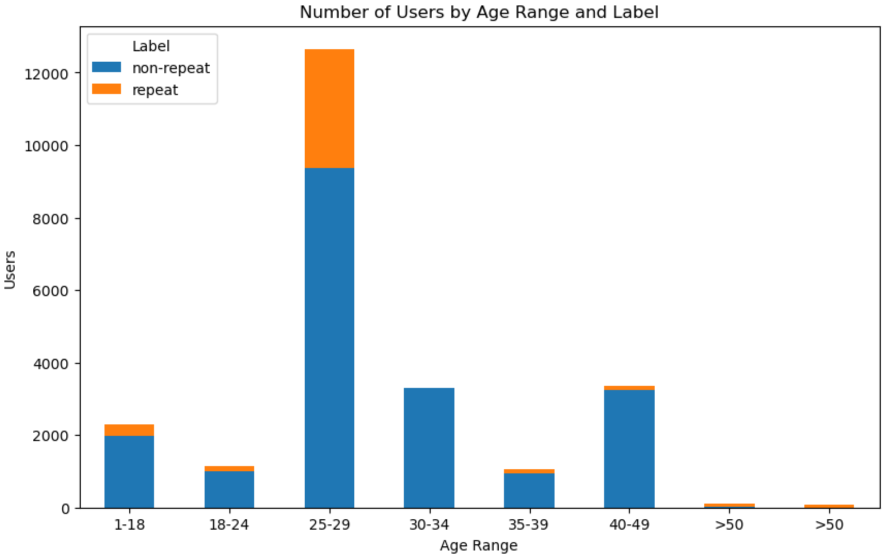
1. Predict the probability of the given user becoming a repeat buyer of the given merchant in the future
2. To find the most important factor to predict repeat buyers

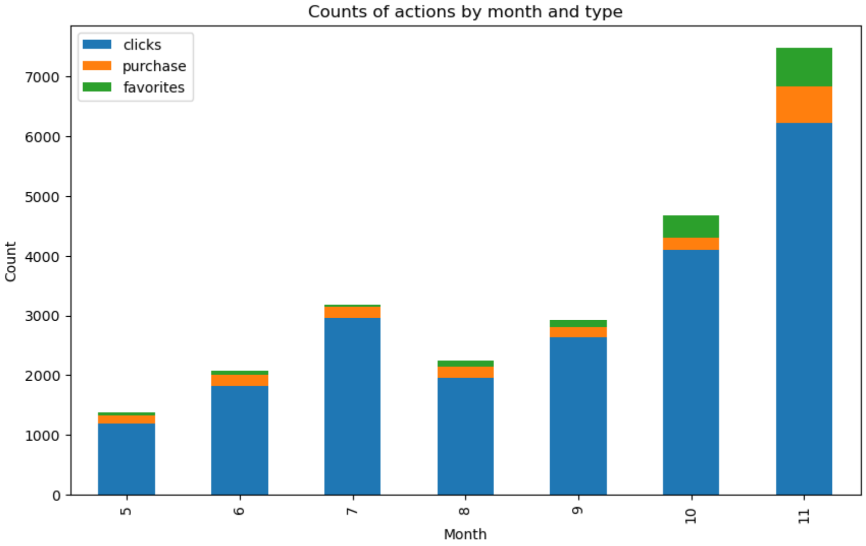
Dataset

- The merged dataset from the given data profiles : user_info, user_log, train_data

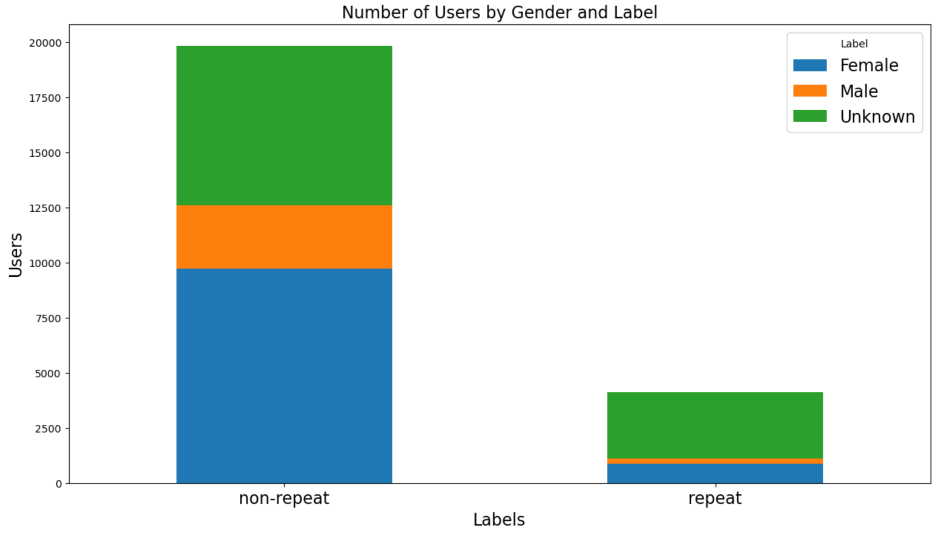
	user_id	merchant_id	label	item_id	cat_id	seller_id	brand_id	time_stamp	action_type	age_range	gender
8291	18306	4733	0	1014866	1397	127	6434	701	0	3	2
19281	99459	2032	0	162933	1389	4963	1991	1002	0	6	0
2893	18306	1710	1	66013	1577	3877	3213	1005	0	3	2
22182	238467	4966	0	935412	1349	1629	2292	821	3	6	1
10805	149634	742	0	236773	1505	416	4014	812	0	6	0

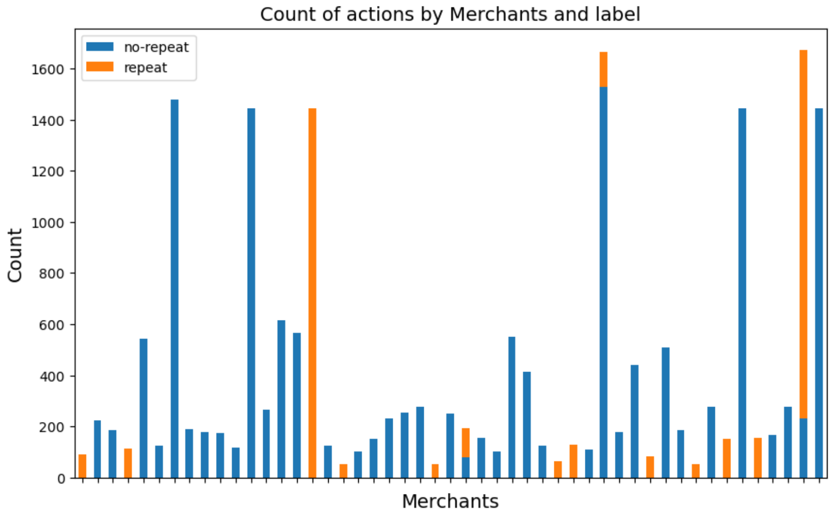
Data Visualization





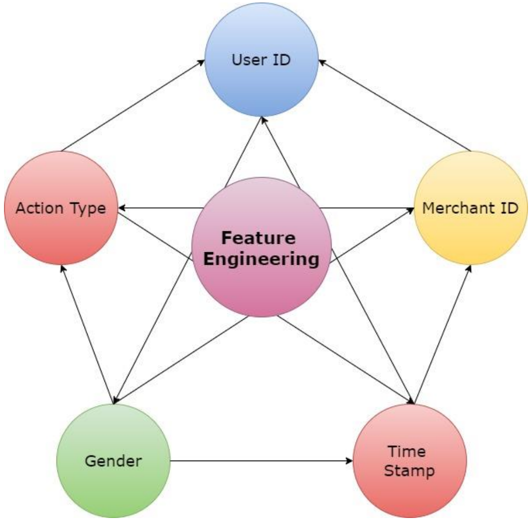
Data Visualization





Feature Engineering

- Average User Age for each Category
- Purchase Average Time
- Purchase Ratio
- Purchase Frequency
- Average User Age for each Merchant
- Ratio of add-to-cart actions to clicks for each Merchant
- Number of distinct brands a user has interacted with for each Merchant
- Number of distinct categories a user has interacted with for each Merchant



Feature Ranking

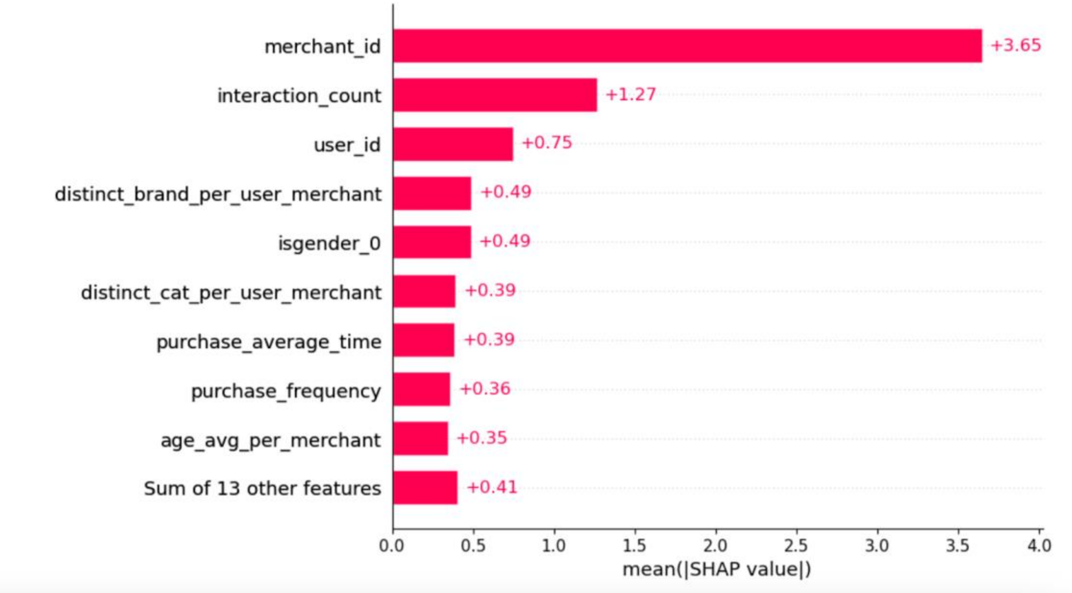
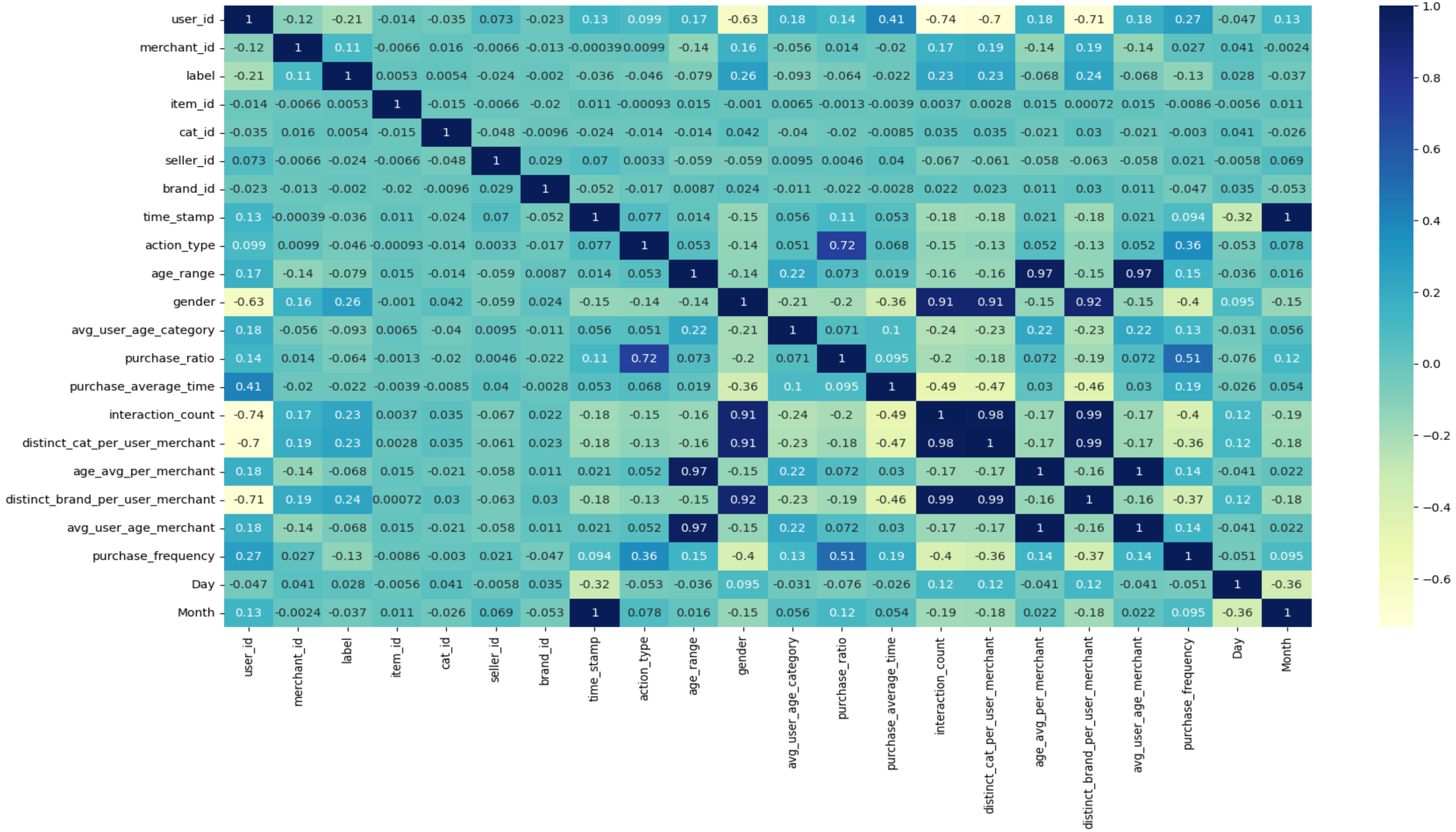
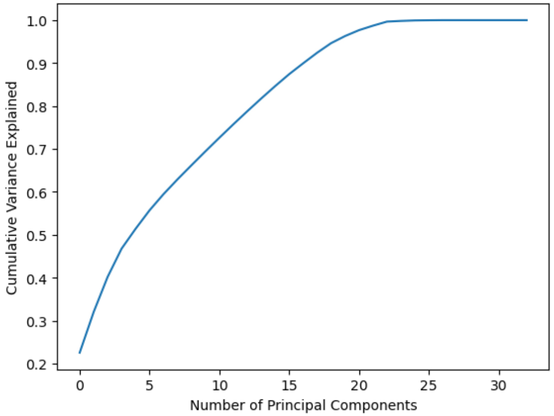


Figure 1: Important features based on SHAP


user_id 1 -0.12 -0.21 -0.014 -0.035 0.073 -0.023 0.13 0.099 0.17 -0.63 0.18 0.14 0.41 -0.74 -0.7 0.18 -0.71 0.18 0.27 -0.047 0.13
merchant_id -0.12 1 0.11 -0.0066 0.016 -0.0066 -0.013 -0.00039 0.0099 -0.14 0.16 -0.056 0.014 -0.02 0.17 0.19 -0.14 0.19 -0.14 0.027 0.041 -0.0024
label -0.21 0.11 1 0.0053 0.0054 -0.024 -0.002 -0.036 -0.046 -0.079 0.26 -0.093 -0.064 -0.022 0.23 0.23 -0.068 0.24 -0.068 -0.13 0.028 -0.037
item_id -0.014 -0.0066 0.0053 1 -0.015 -0.0066 -0.02 0.011 -0.00093 0.015 -0.001 0.0065 -0.0013 -0.0039 0.0037 0.0028 0.015 0.00072 0.015 -0.0086 -0.0056 0.011
cat_id -0.035 0.016 0.0054 -0.015 1 -0.048 -0.0096 -0.024 -0.014 -0.014 0.042 -0.04 -0.02 -0.0085 0.035 0.035 -0.021 0.03 -0.021 -0.003 0.041 -0.026
seller_id 0.073 -0.0066 -0.024 -0.0066 -0.048 1 0.029 0.07 0.0033 -0.059 -0.059 0.0095 0.0046 0.04 -0.067 -0.061 -0.058 -0.063 -0.058 0.021 -0.0058 0.069
brand_id -0.023 -0.013 -0.002 -0.02 -0.0096 0.029 1 -0.052 -0.017 0.0087 0.024 -0.011 -0.022 -0.0028 0.022 0.023 0.011 0.03 0.011 -0.047 0.035 -0.053
time_stamp 0.13 -0.00039 -0.036 0.011 -0.024 0.07 -0.052 1 0.077 0.014 -0.15 0.056 0.11 0.053 -0.18 -0.18 0.021 -0.18 0.021 0.094 -0.32 1
action_type 0.099 0.0099 -0.046 -0.00093 -0.014 0.0033 -0.017 0.077 1 0.053 -0.14 0.051 0.72 0.068 -0.15 -0.13 0.052 -0.13 0.052 0.36 -0.053 0.078
age_range 0.17 -0.14 -0.079 0.015 -0.014 -0.059 0.0087 0.014 0.053 1 -0.14 0.22 0.073 0.019 -0.16 -0.16 0.97 -0.15 0.97 0.15 -0.036 0.016
gender -0.63 0.16 0.26 -0.001 0.042 -0.059 0.024 -0.15 -0.14 -0.14 1 -0.21 -0.2 -0.36 0.91 0.91 -0.15 0.92 -0.15 -0.4 0.095 -0.15
avg_user_age_category 0.18 -0.056 -0.093 0.0065 -0.04 0.0095 -0.011 0.056 0.051 0.22 -0.21 1 0.071 0.1 -0.24 -0.23 0.22 -0.23 0.22 0.13 -0.031 0.056
purchase_ratio 0.14 0.014 -0.064 -0.0013 -0.02 0.0046 -0.022 0.11 0.72 0.073 -0.2 0.071 1 0.095 -0.2 -0.18 0.072 -0.19 0.072 0.51 -0.076 0.12
purchase_average_time 0.41 -0.02 -0.022 -0.0039 -0.0085 0.04 -0.0028 0.053 0.068 0.019 -0.36 0.1 0.095 1 -0.49 -0.47 0.03 -0.46 0.03 0.19 -0.026 0.054
interaction_count -0.74 0.17 0.23 0.0037 0.035 -0.067 0.022 -0.18 -0.15 -0.16 0.91 -0.24 -0.2 -0.49 1 0.98 -0.17 0.99 -0.17 -0.4 0.12 -0.19
distinct_cat_per_user_merchant -0.7 0.19 0.23 0.0028 0.035 -0.061 0.023 -0.18 -0.13 -0.16 0.91 -0.23 -0.18 -0.47 0.98 1 -0.17 0.99 -0.17 -0.36 0.12 -0.18
age_avg_per_merchant 0.18 -0.14 -0.068 0.015 -0.021 -0.058 0.011 0.021 0.052 0.97 -0.15 0.22 0.072 0.03 -0.17 -0.17 1 -0.16 1 0.14 -0.041 0.022
distinct_brand_per_user_merchant -0.71 0.19 0.24 0.00072 0.03 -0.063 0.03 -0.18 -0.13 -0.15 0.92 -0.23 -0.19 -0.46 0.99 0.99 -0.16 1 -0.16 -0.37 0.12 -0.18
avg_user_age_merchant 0.18 -0.14 -0.068 0.015 -0.021 -0.058 0.011 0.021 0.052 0.97 -0.15 0.22 0.072 0.03 -0.17 -0.17 1 -0.16 1 0.14 -0.041 0.022
purchase_frequency 0.27 0.027 -0.13 -0.0086 -0.003 0.021 -0.047 0.094 0.36 0.15 -0.4 0.13 0.51 0.19 -0.4 -0.36 0.14 -0.37 0.14 1 -0.051 0.095
Day -0.047 0.041 0.028 -0.0056 0.041 -0.0058 0.035 -0.32 -0.053 -0.036 0.095 -0.031 -0.076 -0.026 0.12 0.12 -0.041 0.12 -0.041 -0.051 1 -0.36
Month 0.13 -0.0024 -0.037 0.011 -0.026 0.069 -0.053 1 0.078 0.016 -0.15 0.056 0.12 0.054 -0.19 -0.18 0.022 -0.18 0.022 0.095 -0.36 1
user_id merchant_id label item_id cat_id seller_id brand_id time_stamp action_type age_range gender avg_user_age_category purchase_ratio purchase_average_time interaction_count distinct_cat_per_user_merchant age_avg_per_merchant distinct_brand_per_user_merchant avg_user_age_merchant purchase_frequency Day Month
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

Principal Component Analysis (PCA)

- We performed principal component analysis (PCA) on the data to reduce its dimensionality.
- We determined the optimal number of components by selecting the smallest number that captured at least 95% of the variance in the original data. 20 components in our case.
- We then used the PCA-transformed data to train the Bayesian Gaussian classifier . We found that this approach resulted in a much better accuracy of 75% from 63%.



Prediction Models

Splitting of Data

- Training Set : 80% of data
- Validation Set : 20% of Training Set
- Testing Set : 20% of data

Trained on the following models

- Bayes classifier
- Random Forest
- KNN
- Neural Networks

Hyperparameter Tuning

- Hyperparameter tuning involves selecting the optimal values for model parameters that are set before training to improve the model's performance. Such as loss functions, learning rate, activation functions and optimizers.

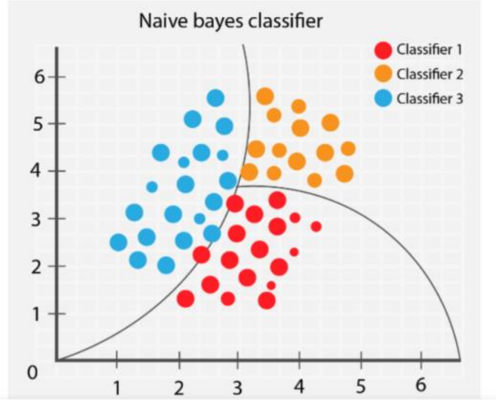
- By tuning hyperparameters, we can improve a model's performance on unseen data and avoid overfitting or underfitting.

- A validation set is used to evaluate the model during hyperparameter tuning, which helps in selecting the best set of hyperparameters that generalize well to new data.

Bayes classifier:

- Bayes' theorem is a fundamental principle in probability theory that describes the probability of an event based on prior knowledge or information.

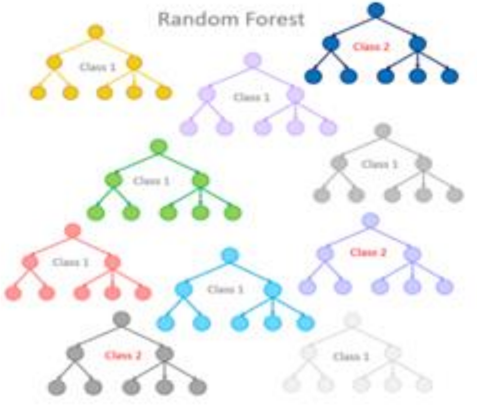
Model	Accuracy	F1 Score
Bayes	0.75	0.36



Random forest:

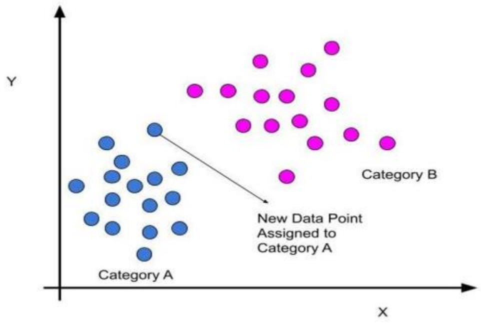
- Different classifiers overfit the data in a different way, and through voting those differences are **averaged out.**

Model	Accuracy	F1 Score
RFC	0.86	0.58



KNN:

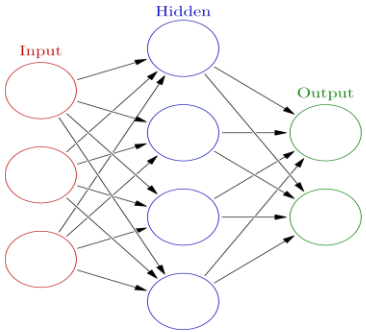
- KNN with k=1 and distance metric 'manhatten' performed best.
- F1-score was low due to the data being unbalanced.
- After **upsampling** on training data the metrics improved.



Model	Accuracy	F1 Score
KNN	0.84	0.56

Neural Network

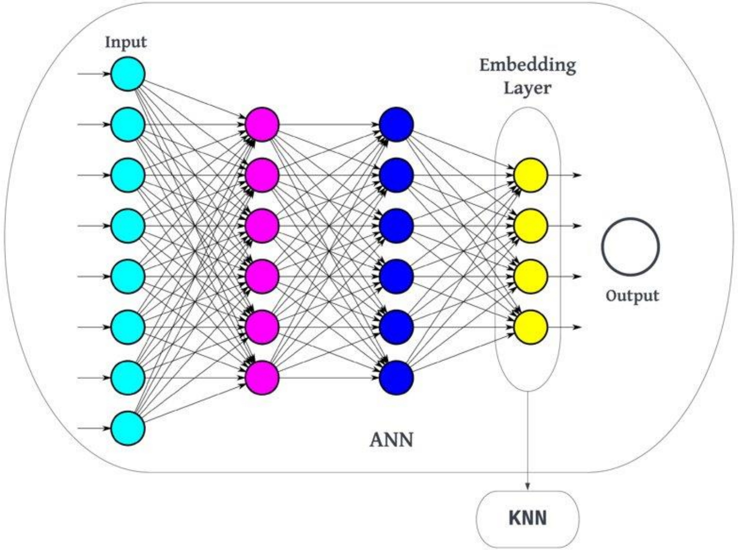
- The optimal model has 4 dense layers with 'relu' activation and an output layer with 'sigmoid' activation function.
- Binary Cross Entropy loss showed the best accuracy with 'adam' optimizer.
- **Dropout** layers were used for regularization.



Model	Accuracy	F1 Score
ANN	0.94	0.85

ANN combined with KNN

- The ANN model was combined with KNN.
- ANN was trained on the data and the penultimate layer(embedding) was then fed to the KNN.
- ANNs and KNN have complementary strengths and weaknesses. ANNs are good at learning complex non-linear patterns in the data, while KNN is good at capturing the local structure of the data.

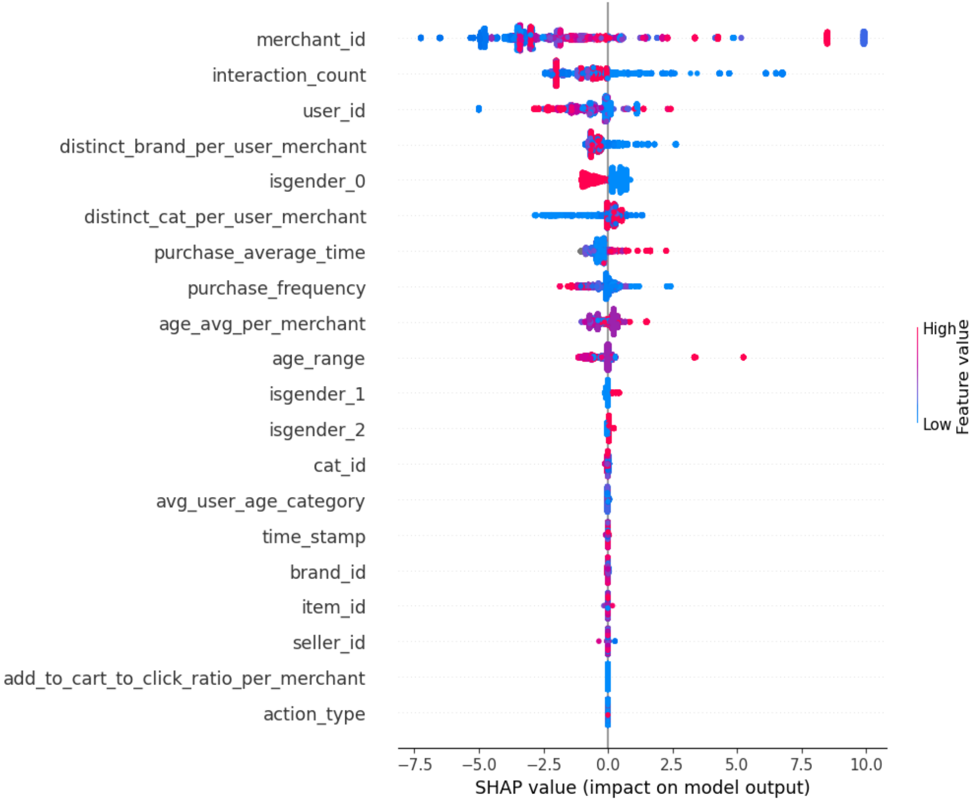


Model	Accuracy	F1 Score
ANN + KNN	0.96	0.90

Models Performance

Model	Accuracy	F1 Score
Bayes	0.75	0.36
RFC	0.86	0.68
KNN	0.84	0.56
ANN	0.94	0.85
ANN + KNN	**0.96**	**0.90**

Results


merchant_id
interaction_count
user_id
distinct_brand_per_user_merchant
isgender_0
distinct_cat_per_user_merchant
purchase_average_time
purchase_frequency
age_avg_per_merchant
age_range
isgender_1
isgender_2
cat_id
avg_user_age_category
time_stamp
brand_id
item_id
seller_id
add_to_cart_to_click_ratio_per_merchant
action_type
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
SHAP value (impact on model output)
High
Low
Feature value

Recommendations

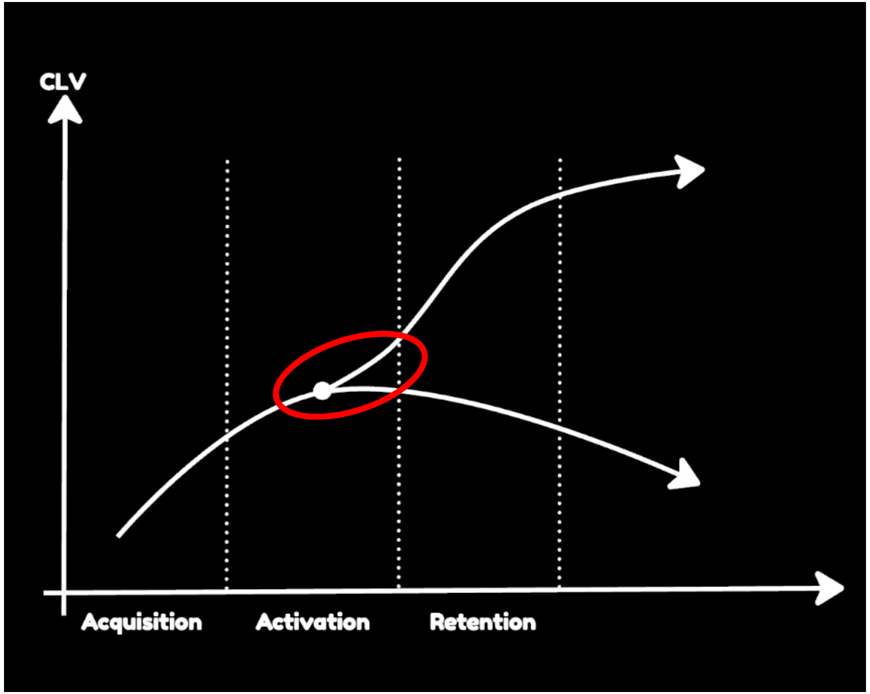
- **Improve user engagement**
  - businesses can consider offering personalized recommendations based on the user's browsing and purchase history.

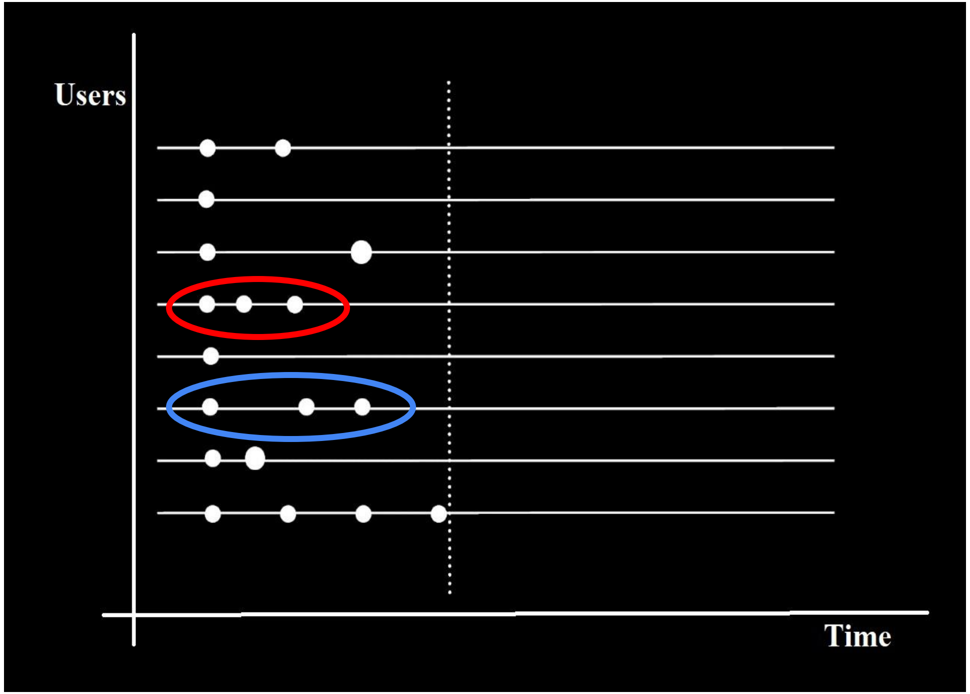
- **Segments users by gender and age range**
  - businesses can tailor their offerings and marketing strategies to specific groups

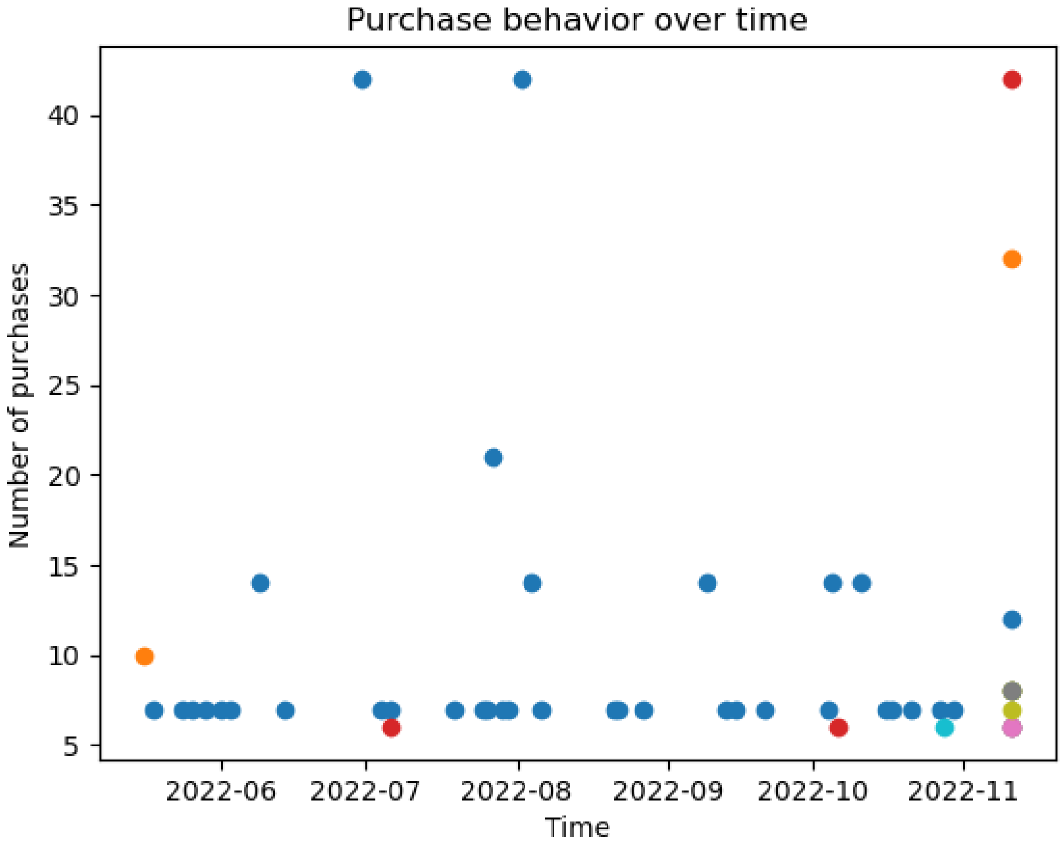
- **Increase purchase frequency**
  - businesses can offer promotions or incentives to encourage users to make repeat purchases.

- **Personalize marketing strategies**
  - businesses can personalize their marketing strategies to target specific users with products or promotions that are most likely to appeal to them.

Recommendations






Purchase behavior over time
Number of purchases
Time
40
35
30
25
20
15
10
5
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11

Conclusion

- The methods predict whether a user would be a repeat customer or not by training different models on a dataset containing user interactions with merchants.
- The model ANN combined KNN achieved the highest accuracy and F1-score.
- The most important features for predicting repeat customers were found to be interaction_count, merchant_id, and gender.

Overall, the study provides valuable insights for businesses looking to improve customer retention.