Final Report

Evaluating the demand for a cat cafe in Dhaka

CSE 445.10
(Machine Learning)

Group 2

Prepared by:
Rashedul Islam Tono (2022154642)
Wahidun Akter (2022188642)
Ashfaqur Rahman Chowdhury (2111743042)
Shakhawat Hossain (2111555042)

May 16, 2024

Table of Contents

Overview

This project aims to assess the customer demand for a potential cat cafe in Dhaka, Bangladesh. Through surveys and subsequent data analysis, we aimed to create a model that predicts the feasibility of such a business based on some demographic info and user preferences such as location, duration, spending nature, etc.

Hypothesis

Based on general assumptions, we came up with three hypotheses for this project-

H1: Customers who love cats will prefer to stay longer at the cafe.

H2: Customers in the younger age demographic are more encouraged to visit the cat cafe just because they love cats.

H3: Female customers are more likely to be strong cat lovers, and stay longer at the cat cafe compared to Male customers.

04

Dataset description

```
df5.head()
```

	Age	Gender	Occupation	Cat Lover	Dislike Reason	Cafe Duration	Encouragement	Place	Spend	Additional
0	20-29	Male	Student	Strongly agree, I love cats!	No preference	1hr	Discounts on menu items	Uttara	Less than 250 BDT	Cat accessories
1	20-29	Male	Student	Agree, I like cats	Hygiene Concerns	No preference	Discounts on menu items	Bashundhara R/A	250-450 BDT	No entrance fees or cat products
2	20-29	Male	Student	Strongly agree, I love cats!	No preference	1hr	New and innovative menu items	Uttara	250-450 BDT	No entrance fees or cat products
3	20-29	Male	Student	Strongly agree, I love cats!	No preference	1hr	All of the above	Gulshan	250-450 BDT	No preference
4	20-29	Male	Student	Agree, I like cats	No preference	1hr	Exclusive offers for regular customers	Gulshan	250-450 BDT	No entrance fees or cat products

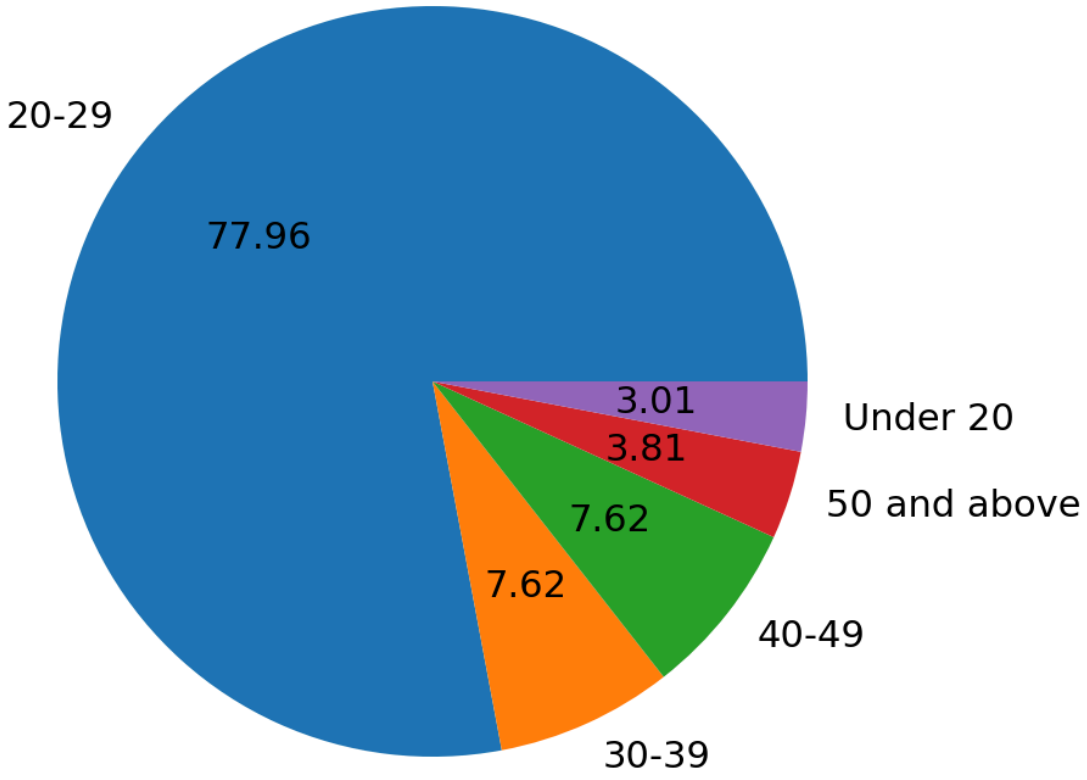
```
df5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 499 entries, 0 to 507
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             499 non-null    object
 1   Gender          499 non-null    object
 2   Occupation      499 non-null    object
 3   Cat Lover       499 non-null    object
 4   Dislike Reason  499 non-null    object
 5   Cafe Duration   499 non-null    object
 6   Encouragement   499 non-null    object
 7   Place           499 non-null    object
 8   Spend           499 non-null    object
 9   Additional      499 non-null    object
dtypes: object(10)
memory usage: 42.9+ KB
```

Our initial dataset had 508 responses, after data cleaning it became 499. The 'Age','Gender' and 'Occupation' columns represent demographic data of the participants, and rest of the columns/features represent their preference for a cat cafe in Dhaka.

Data Analysis

Age demographic



After analyzing the dataset, we can see that the majority of the participants were within the age group of 20-29, meaning our dataset is biased towards younger audience.

Gender demographic

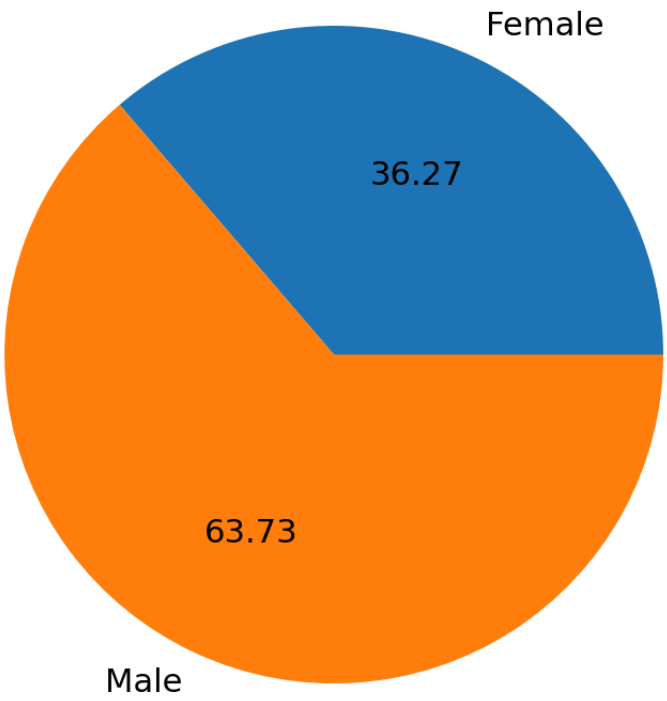


Fig. Gender ratio in the overall dataset

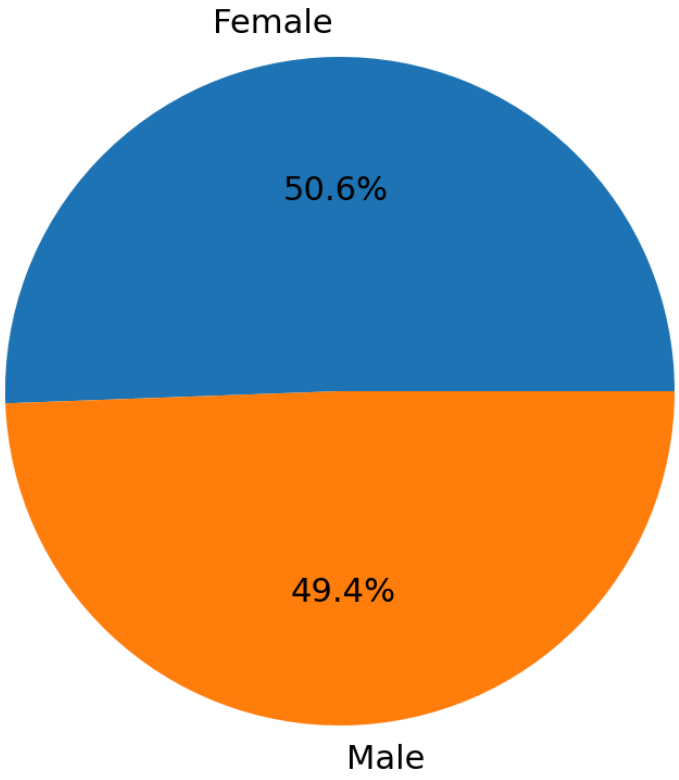


Fig. Percentage of Male vs Female who strongly love cats

As for the gender, although most participants were Male, we can also see that the majority of the Females strongly love cats, which further bolsters our 3rd Hypothesis.

One Hot encoding

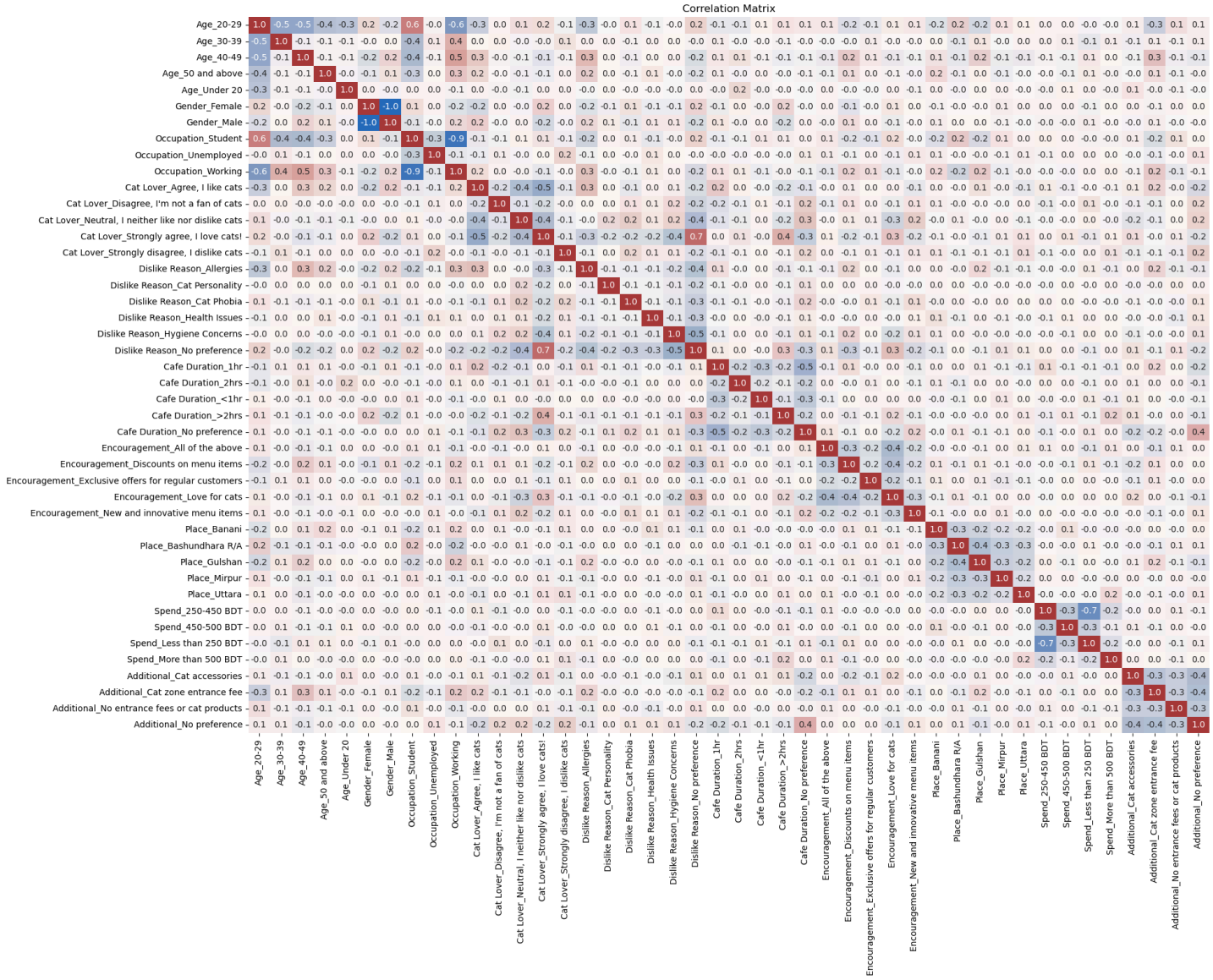
```
from sklearn.preprocessing import OneHotEncoder

categorical_columns = df5.select_dtypes(include=['object']).columns.tolist()

encoder = OneHotEncoder(sparse_output=False)

one_hot_encoded = encoder.fit_transform(df5[categorical_columns])

one_hot_df = pd.DataFrame(one_hot_encoded, columns=encoder.get_feature_names_out(categorical_columns))

df7 = pd.concat([df5, one_hot_df], axis=1)

df7 = df7.drop(categorical_columns, axis=1)

df7.head()
```

	Age_20-29	Age_30-39	Age_40-49	Age_50 and above	Age_Under 20	Gender_Female	Gender_Male	Occupation_Student	Occupation_Unemployed	Occupation_Working	...
0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...
1	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...
2	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...
3	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...
4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...

5 rows × 44 columns

Since we wanted to visualize the correlation between each of the categorical options under each column/feature, we chose One Hot Encoding, which groups all legal combinations of categorical values as their own separate column/feature using binary encoding. This will help us perform a detailed correlation analysis between each of the categorical options.

Correlation Matrix



From this correlation matrix, we can see that-

✅ H1: "Cat lover: Strongly Agree" is highly correlated (0.4) with "Cafe duration: >2hrs", which proves that customers who loves cat will prefer to stay longer at the cat cafe.

❌ H2: "Age: 20-29" is not just solely correlated with "Encouragement: Love for cats", rather other encouragement factors such as "New and Innovative Menu items" are also equally correlated (0.1). So this hypothesis does not stand.

✅ H3: "Gender: Female" is highly correlated (0.2) with both "Cat lover: Strongly agree" and "Cafe duration: >2hrs", which proves that female customers are strong cat lovers and they will prefer to spend a significant amount of time at the cat cafe.

Additional Observations from Correlation Matrix

1. Age 40 and above have allergy issues as their main reason for disliking cats. Moreover, it's the most chosen reason for disliking cats among all the participants.

2. Age 40 and above, mainly the working population preferred Banani and Gulshan for the Cafe, whereas the majority of the younger demographics who are students preferred Bashundhara R/A.

3. The majority of the participants who were neutral, or disliked or strongly disliked cats had no preference for the cafe staying duration, nor any additional features, and were mostly interested in innovative menu items rather than cats, indicating they are not that interested in the concept of a Cat Cafe in Dhaka.

Feature Selection

Determining Target Function

Our main objective of this model is to predict the feasibility of opening a cat cafe in Dhaka or not, and to train such model we need a target column/target function which contains whether the customer is interested in this Cat Cafe or not (Y/N).
Although we don't have such target column in our dataset, from our correlation analysis we can clearly see that majority of the participants who dislike cats/neutral had either no preference, or <1hr of preferred time to stay in the cat cafe.
Based on this observation, we can group the dataset into two sets, one who are interested and another who aren't that interested in the Cat Cafe.

```
df8['Interest(Target)'] = 1.0
```

```
def interest_logic1(row):
  return 0 if (row['Cat Lover_Disagree, I\'m not a fan of cats'] == 1 and
               (row['Cafe Duration_No preference'] == 1 or row['Cafe Duration_<1hr'] == 1) and
               row['Additional_No preference'] == 1) else 1

# Add a new column named "Interest(Target)" with the logic applied
df8['Interest(Target)'] = df8.apply(interest_logic1, axis=1)
```

```
def interest_logic2(row):
  return 0 if (row['Cat Lover_Strongly disagree, I dislike cats'] == 1 and
               (row['Cafe Duration_No preference'] == 1 or row['Cafe Duration_<1hr'] == 1) and
               row['Additional_No preference'] == 1) else 1

# Add a new column named "Interest(Target)" with the logic applied
df8['Interest(Target)'] = df8.apply(interest_logic2, axis=1)
```

```
def interest_logic3(row):
  return 0 if (row['Cat Lover_Neutral, I neither like nor dislike cats'] == 1 and
               (row['Cafe Duration_No preference'] == 1 or row['Cafe Duration_<1hr'] == 1) and
               row['Additional_No preference'] == 1) else 1

# Add a new column named "Interest(Target)" with the logic applied
df8['Interest(Target)'] = df8.apply(interest_logic3, axis=1)
```

Dimensionality Reduction

From our Correlation Matrix and verified hypotheses, we came to the conclusion that 'Age, Occupation, Dislike Reason, Encouragement, Additional'- features do not have a high correlation value with the target function. Thus, we omitted these features/columns from our model training.

```
#relevant features: gender, cot lovers, duration, location, spend
#dropped columns: age, occupation, dislike reason, encouragement, additional
df8.drop(columns=df8.columns[[0,1,2,3,4,7,8,9,15,16,17,18,19,20,26,27,28,29,30,40,41,42,43]], inplace=True)
```

Model performance and Evaluation

Train/Test split

We split our dataset into 80-20 ratio, 80% of which is used for training, and 20% for testing the performance of the models.

Linear Regression

```
R2 score:  -0.4309976398149029
MSE score:  0.06255435943603516
RMSE score:  0.2501086952427587
MAE score:  0.16771484375
```

Here, although our MSE and RMSE scores are not inherently bad, a negative R^2 score of -0.43 indicates that the linear regression model actually increases the variability of the target variable compared to just using the mean of the target variable. This is a strong signal that linear regression is unsuitable for this binary classification problem.

Logistic Regression

```
Accuracy: 0.92
Precision: 0.9438202247191011
Recall: 0.9655172413793104
Confusion Matrix:
 [[ 8  5]
 [ 3 84]]
F1 score:  0.9545454545454545
AUC score:  0.790450928381963
```

Here, our model's accuracy is 92%, which means this model seems to be performing very well at classifying customer interest in our dataset. It has a high accuracy, good precision and recall, and a strong F1-score. While the AUC-ROC score could potentially be improved, it still indicates a clear ability to separate the classes.

KNN

```
Accuracy: 0.92
Precision: 0.9438202247191011
Recall: 0.9655172413793104
Confusion Matrix:
 [[ 8  5]
 [ 3 84]]
F1 score:  0.9545454545454545
AUC score:  0.790450928381963
```

Here, we set the K value to 5, and it performed nearly identical as our Logistic Regression model.

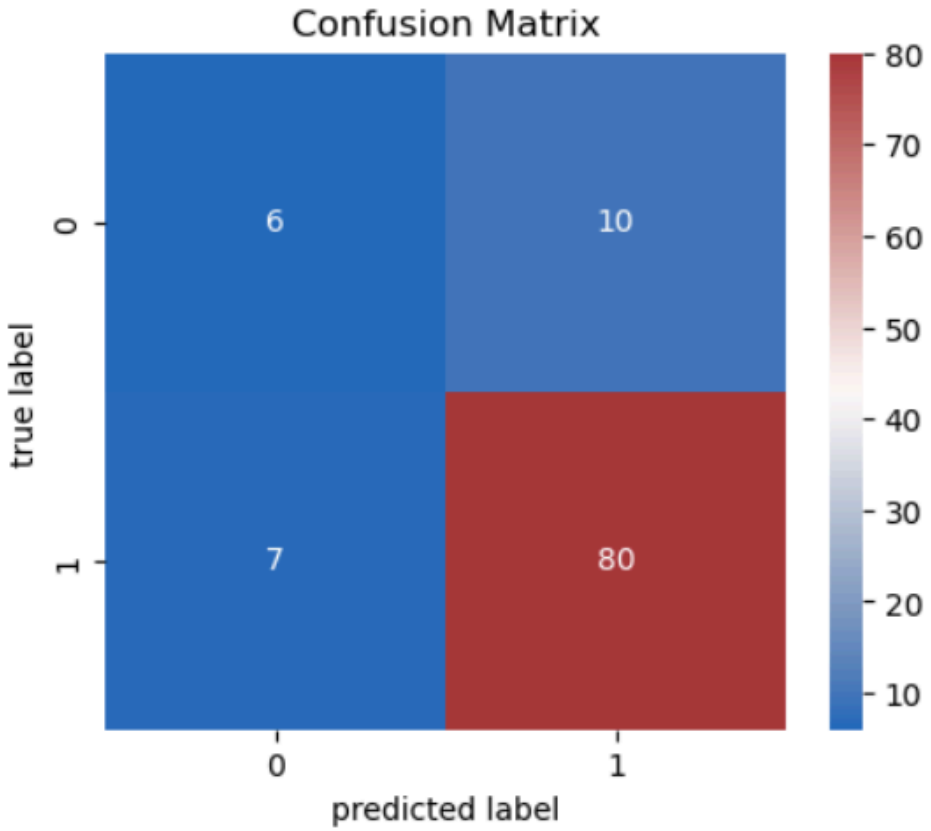
Adding more unbiased data to the model

Since our data was biased towards a certain age demographic, we introduced a few unbiased data to our dataset and retrained the KNN model. This time the accuracy value dropped to 83%

```
Accuracy: 0.8349514563106796
Precision: 0.8888888888888888
Recall: 0.9195402298850575
Confusion Matrix:
 [[ 6 10]
 [ 7 80]]
F1 score:  0.903954802259887
AUC score:  0.6472701149425287
```

Discussion

Judging by our final model's Confusion Matrix, out of 103 test data, it correctly predicted 80 users who are interested in the cat cafe concept, incorrectly classified 7 interested users as not interested (False Negative), and incorrectly classified 10 non-interested users as interested (False Positive).



Therefore, the high True Positive rate (80) and low False Negative rate (7) indicate that opening a cat cafe in Dhaka has the potential to be a successful business endeavor.