August 2022

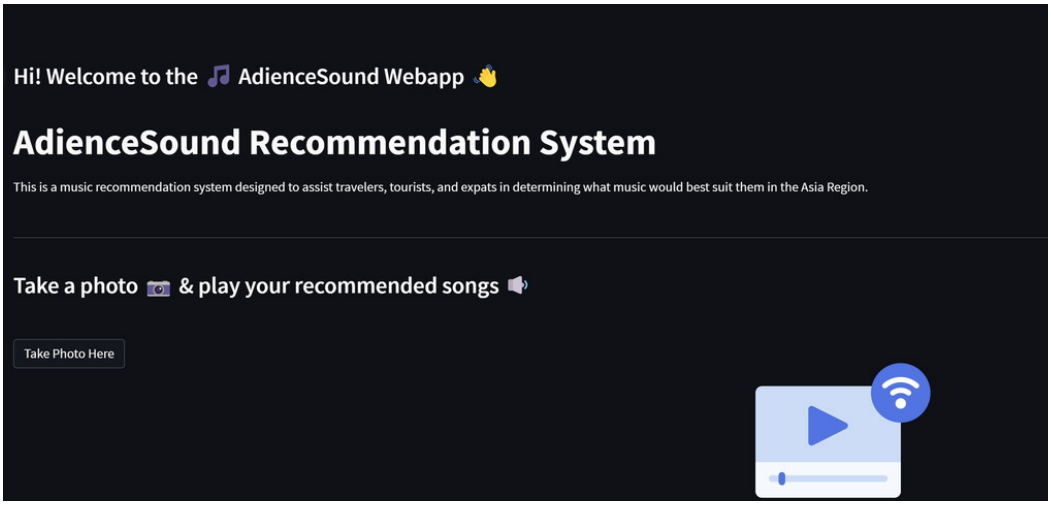
Capstone Project

ADIENCESOUND TECHNICAL DOCUMENTATION

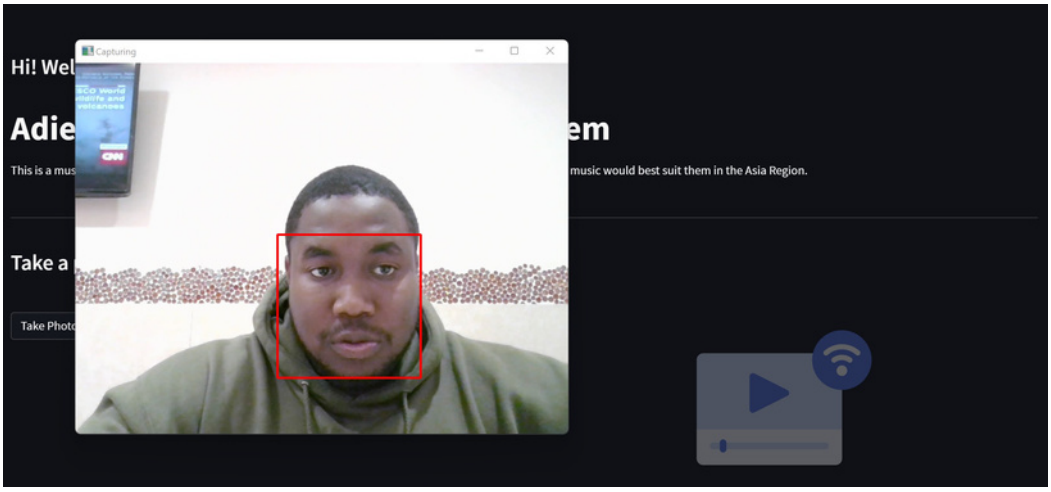
Prepared by:
Clifton Lee

ANNEX: PRODUCT GUI (WEB APP)

VIEW 1



VIEW 2



VIEW 3

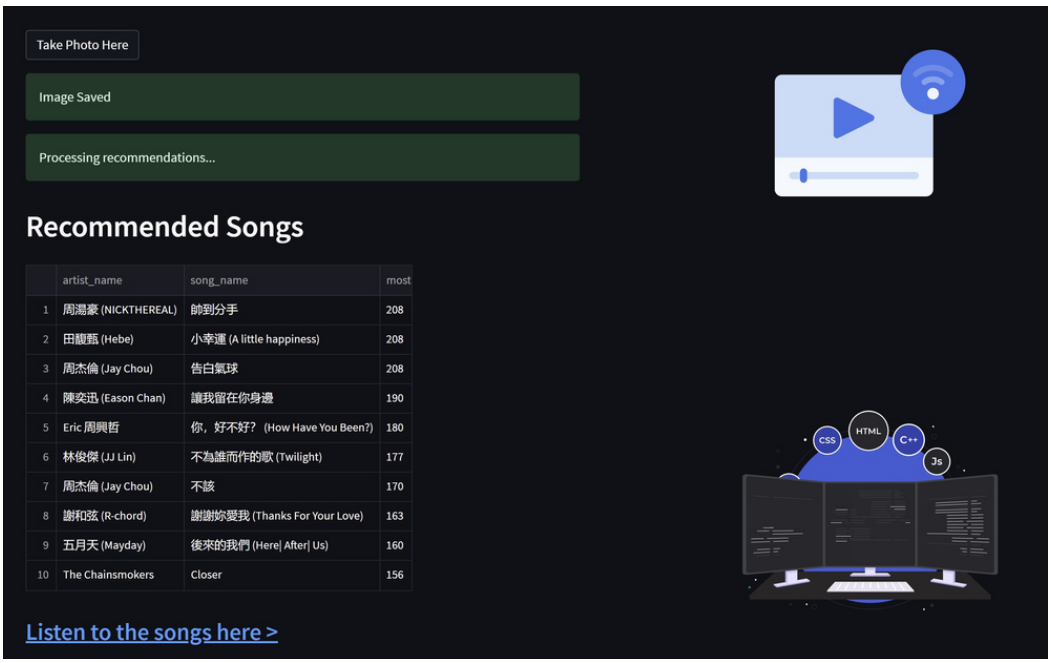


TABLE OF CONTENTS

PURPOSE OF THIS DOCUMENT

This document provides guidance and is intended to aid in the basic understanding of the technical design and functionality of the product.
In order to set up the test product on another web server, both technical and product overviews are provided.

PRODUCT OVERVIEW

The AdienceSound Recommendation System is a collaborative filtering recommendation system based on a model that uses facial recognition to fine-tune each individual's music selection. This is delivered via a web app with a gamification vibe. Views of the app can be found in the Annex section.

TECHNICAL REQUIREMENTS

To test and use the AdienceSound Recommendation System, the following features must be installed:

1. Python V3.8 or above
2. Python packages & modules:
    a. Tensorflow V2 or above
    b. Keras
    c. Numpy V1.2 or above
    d. Pandas V1.2 or above
    e. Matplotlib V3.3
    f. Seaborn
    g. OpenCV V4
    h. Pillow V3
    i. Sklearn
    j. Dlib
    k. Face_recognition
    l. Deepface

DATA DESCRIPTION

The AdienceSound recommendation system was created by the combination of two distinct kinds of datasets.

Sources:

1. For unstructured data, the Adience dataset was used (https://www.kaggle.com/ttungl/adience-benchmark-gender-and-age-classification).
2. For structured data, the KKBOX dataset was used, which is also accessible on the Kaggle site (https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data).

The following is a preliminary discussion of the many forms of data:

1. In the case of unstructured data, The Adience dataset comprises 26,580 images of 2,284 participants with a binary gender label and one label from eight distinct age groups, divided into five splits. The guiding philosophy of this dataset's collection aims to capture photos as closely as possible to real-world situations, including all variances in look, position, lighting condition, and image quality, to mention a few. I've utilized this data for training a model capable of determining the gender and age of a user based on their input. This result was used to assist in making music recommendations.
2. In the case of structured data, The KKBOX dataset comprises approximately 10 million recordings of user music selections, as well as demographic information about the users. KKBOX.com is Asia's most extensive music streaming service, with over 30 million songs. I utilized their data to decipher their users' past choices. This information got me access to the music selection preferences and song ratings of over 10 million consumers.

Dimensions of the Data:

1. Dataset on Adience:
    a. 26,580 images
    b. 2,284 individuals
    c. 8 age groupings
2. Dataset KKBOX:
    a. Table of User Song Selections: 9,934,208 entries with six fields.
    b. Demographic Information on Users: 34,403 records and 7 fields
    c. Songs Table: 2,296,320 rows and seven columns.

METHODOLOGY TECHNIQUE OVERVIEW

The CRISP-DM methodology was used to develop this product. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a time-tested method for guiding data mining efforts.
This methodology consists of six stages. This documentation will include descriptions of typical project phases, tasks associated with each phase, and an explanation of the relationships between these tasks at each stage:

Figure 1. The data mining life cycle

STAGE 1: BUSINESS UNDERSTANDING

At this stage, I asked, "What are the business reasons for the data mining effort?" This aided me in ensuring that the focus should be streamlined and that everyone is on the same page before expending valuable resources.
Following that, I gathered background information on Asia's leading streaming platform (KKBOX) as well as listeners' problems. Data was also gathered to aid in the identification of individuals using images of their faces (Adience dataset).
I then documented the business objectives stated in the initial proposal in order to determine the success of data mining from a business standpoint.
I created a plan for the data mining project. The questions I've posed thus far, as well as the business and data mining objectives I've established, served as the foundation for this road map.

STAGE 2: DATA UNDERSTANDING

The data understanding phase of CRISP-DM entailed taking a closer look at the data available for mining. This step was critical in avoiding unexpected problems during the next phase, data preparation, which was the longest part of the project.
The data understanding stage involved accessing the data and exploring it using tables and graphics. This allowed me to assess the data's quality and comprehend the limitations of my product.

Here are some data analysis snippets from each dataset exploration:

Adience dataset

Here is some metadata :

Columns	*Data types*
user_id	object
original_image	object
face_id	int64
age	object
gender	object
x	int64
y	int64
dx	int64
dy	int64
tilt_ang	int64
fiducial_yaw_angle	int64
fiducial_score	int64

dtype: object

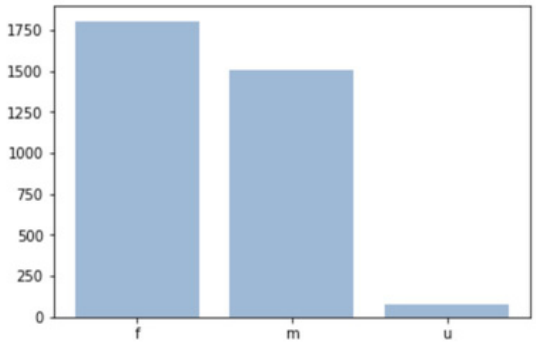
The first 10 rows of the unstructured dataset metadata:

	user_id	original_image	face_id	age	gender	x	y	dx	dy	tilt_ang	fiducial_yaw_angle	fiducial_score
0	113445054@N07	11763777465_11d01c34ce_o.jpg	1322	(25, 32)	m	1102	296	357	357	-15	0	59
1	113445054@N07	11763777465_11d01c34ce_o.jpg	1323	(25, 32)	f	1713	580	325	325	-5	0	118
2	113445054@N07	11763777465_11d01c34ce_o.jpg	1324	(15, 20)	f	1437	664	306	306	5	0	109
3	113445054@N07	11764005785_f21921aea6_o.jpg	1325	(25, 32)	f	978	229	803	803	-20	-45	16
4	113445054@N07	11763728674_a41d99f71e_o.jpg	1326	(25, 32)	m	1745	910	242	242	-10	0	55
5	113445054@N07	11764019623_8ffb8ff4f5_o.jpg	1327	(25, 32)	f	1294	752	1013	1013	-10	30	110
6	113445054@N07	11764019623_8ffb8ff4f5_o.jpg	1325	(25, 32)	f	798	583	943	943	-10	15	57
7	113445054@N07	11764019623_8ffb8ff4f5_o.jpg	1328	(25, 32)	f	2632	1069	243	242	15	15	23
8	113445054@N07	11763616596_db19dbce85_o.jpg	1329	34	m	803	854	612	612	5	0	20
9	113445054@N07	11763616596_db19dbce85_o.jpg	1325	(25, 32)	f	1141	1282	503	504	5	0	72

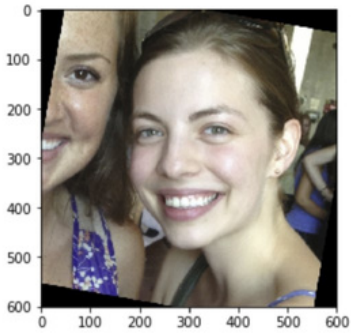
STAGE 2: DATA UNDERSTANDING (Continue)

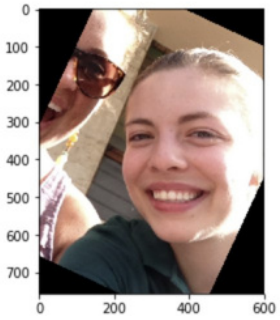
Adience dataset (continue)

Gender Class Distribution:



Examples of what images look like:





STAGE 2: DATA UNDERSTANDING (Continue)

KKBOX Dataset

Column count and data types

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100000 entries, 0 to 99999
Data columns (total 20 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   msno                    100000 non-null  object
 1   song_id                 100000 non-null  object
 2   source_system_tab       99697 non-null   object
 3   source_screen_name      95727 non-null   object
 4   source_type             99805 non-null   object
 5   target                  100000 non-null  int64
 6   song_length             99996 non-null   float64
 7   genre_ids               98498 non-null   object
 8   artist_name             99996 non-null   object
 9   composer                78528 non-null   object
 10  lyricist                59309 non-null   object
 11  language                99996 non-null   float64
 12  name                    99991 non-null   object
 13  isrc                    91475 non-null   object
 14  city                    100000 non-null  int64
 15  bd                      100000 non-null  int64
 16  gender                  61328 non-null   object
 17  registered_via          100000 non-null  int64
 18  registration_init_time  100000 non-null  int64
 19  expiration_date         100000 non-null  int64
dtypes: float64(2), int64(6), object(12)
memory usage: 16.0+ MB
```

First 5 rows' image (too big to capture all columns in image)

	msno	song_id	source_system_tab	source_screen_name
0	FGtllVqz18RPiwJj/edr2gV78zirAiY/9SmYvia+kCg=	BBzumQNXUHKdEBOB7mAJuzok+IJA1c2Ryg/yzTF6tik=	explore	Explore
1	Xumu+NIjS6QYVxDS4/t3SawvJ7viT9hPKXmf0RtLNx8=	bhp/MpSNoqoxOIB+/l8WPqu6jldth4DIpCm3ayXnJqM=	my library	Local playlist more
2	Xumu+NIjS6QYVxDS4/t3SawvJ7viT9hPKXmf0RtLNx8=	JNWfrrC7zNN7BdMpsISKa4Mw+xVJYNnxXh3/Epw7QgY=	my library	Local playlist more
3	Xumu+NIjS6QYVxDS4/t3SawvJ7viT9hPKXmf0RtLNx8=	2A87tzfnJTSWqD7gIZHisolhe4DMdzkbd6LzO1KHjNs=	my library	Local playlist more
4	FGtllVqz18RPiwJj/edr2gV78zirAiY/9SmYvia+kCg=	3qm6XTZ6MOCU11x8FIVbAGH5l5uMkT3/ZalWG1oo2Gc=	explore	Explore

STAGE 3: DATA PREPARATION

The most important and time-consuming aspect of this project was data preparation. In fact, the data preparation stage is estimated to have taken up 50-70% of the project's time and effort.

The following tasks were performed during the data preparation stage:

- Merging data sets and/or records
- Choosing a sample subset of data
- Aggregating records
- Deriving new attributes
- Sorting data for modeling
- Replacing blank or missing values
- Splitting data into training and test sets

Here are some snippets from that stage:

Adience dataset

Encoding of Gender class:
Before

	age	gender	x	y	dx	dy	img_path
0	(25, 32)	f	0	414	1086	1383	AdienceBenchmarkGenderAndAgeClassification/fac...
1	(25, 32)	m	301	105	640	641	AdienceBenchmarkGenderAndAgeClassification/fac...
2	(25, 32)	f	2395	876	771	771	AdienceBenchmarkGenderAndAgeClassification/fac...

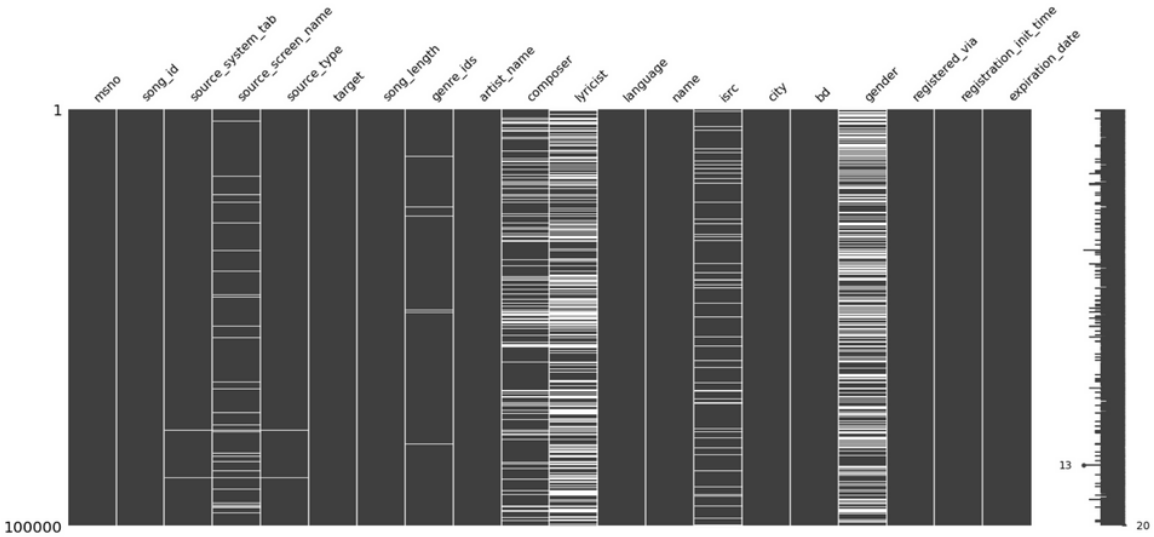
After

	age	gender	x	y	dx	dy	img_path
0	25-32	0	0	414	1086	1383	AdienceBenchmarkGenderAndAgeClassification/fac...
1	25-32	1	301	105	640	641	AdienceBenchmarkGenderAndAgeClassification/fac...
2	25-32	0	2395	876	771	771	AdienceBenchmarkGenderAndAgeClassification/fac...

STAGE 3: DATA PREPARATION (Continue)

KKBOX dataset

Dealing with missing values



Code to fill & replace missing values

```
for i in songs.select_dtypes(include=['object']).columns:
    songs[i][songs[i].isnull()] = 'unknown'
songs = songs.fillna(value=0)
```

STAGE 4: MODELING

This is when all of my hard work began to pay off. The data that I spent time preparing was combined and worked on with some algorithms, and the results began to shed some light on the business problem that was posed during Business Understanding.
This stage was completed in several iterations. I ran several algorithms on the data with the default parameters, then fine-tuned the parameters and returned to the data preparation phase for any necessary manipulations.

Algorithms ran on the Adience dataset:

- Eigenfaces
- Fisherface
- Convolutional Neural Network

Algorithms ran on the KKBox dataset:

- Logistic Regression
- Linear Discriminant Analysis
- K Nearest Neighbors
- Classification and Regression Trees (Decision Trees)
- Gaussian Naive Bayes

STAGE 5: EVALUATION

At this point, I've completed the majority of my data mining project. In the Modeling phase, I also determined that the models created are technically correct and effective according to the data mining success criteria that I defined earlier.
However, before proceeding, I assessed the results of my efforts using the business success criteria established at the start of the project. This was critical to ensuring that the results I obtained could be used by any organization.

STAGE 6: DEPLOYMENT

At this point, I gathered everything and presented the finished product as a Web App with a dark sleek background and gamification icons. This is what users will see, while everything done in the other stages will be in the product's backend. Images of the Web App can be seen the Annex section.

MODEL & EVALUATION

In this section, I will discuss the models chosen for the structured and unstructured datasets, as well as show the evaluation results for each:

Adience dataset (unstructured data): A CNN-based model was used
KKBox dataset (structured data): A classification and Regression Trees based model was used

Here are images displaying the Evaluation:

Adience dataset (*unstructured data)*

Gender Model evaluation

```
Epoch 1/25
382/382 [==============================] - 11s 27ms/step - loss: 0.6899 - accuracy: 0.5414 - val_loss: 0.6925 - val_accuracy: 0.5241
Epoch 2/25
382/382 [==============================] - 10s 26ms/step - loss: 0.6909 - accuracy: 0.5360 - val_loss: 0.6926 - val_accuracy: 0.5241
Epoch 3/25
382/382 [==============================] - 10s 26ms/step - loss: 0.6910 - accuracy: 0.5345 - val_loss: 0.6922 - val_accuracy: 0.5241
Epoch 4/25
382/382 [==============================] - 10s 26ms/step - loss: 0.6905 - accuracy: 0.5370 - val_loss: 0.6927 - val_accuracy: 0.5241
Epoch 5/25
382/382 [==============================] - 10s 26ms/step - loss: 0.6900 - accuracy: 0.5404 - val_loss: 0.6926 - val_accuracy: 0.5241
Epoch 6/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6895 - accuracy: 0.5419 - val_loss: 0.6922 - val_accuracy: 0.5241
Epoch 7/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6901 - accuracy: 0.5402 - val_loss: 0.6925 - val_accuracy: 0.5241
Epoch 8/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6894 - accuracy: 0.5447 - val_loss: 0.6921 - val_accuracy: 0.5241
Epoch 9/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6899 - accuracy: 0.5426 - val_loss: 0.6922 - val_accuracy: 0.5241
Epoch 10/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6909 - accuracy: 0.5340 - val_loss: 0.6930 - val_accuracy: 0.5241
Epoch 11/25
382/382 [==============================] - 10s 27ms/step - loss: 0.6891 - accuracy: 0.5452 - val_loss: 0.6923 - val_accuracy: 0.5241
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

Age Model Summary

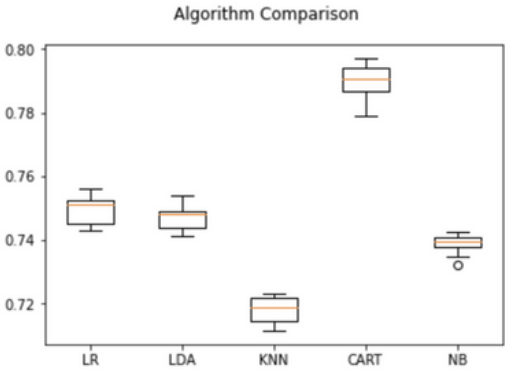
```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_3 (Conv2D)            (None, 56, 56, 96)        14208
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 28, 28, 96)        0
_________________________________________________________________
layer_normalization_3 (Layer (None, 28, 28, 96)        192
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 28, 28, 256)       614656
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 14, 14, 256)       0
_________________________________________________________________
layer_normalization_4 (Layer (None, 14, 14, 256)       512
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 14, 14, 256)       590080
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 7, 7, 256)         0
_________________________________________________________________
layer_normalization_5 (Layer (None, 7, 7, 256)         512
_________________________________________________________________
flatten_1 (Flatten)          (None, 12544)             0
_________________________________________________________________
dense_3 (Dense)              (None, 512)               6423040
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_4 (Dense)              (None, 512)               262656
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_5 (Dense)              (None, 8)                 4104
=================================================================
Total params: 7,909,960
Trainable params: 7,909,960
Non-trainable params: 0
_________________________________________________________________
```

MODEL & EVALUATION

KKBOX dataset (*unstructured data)*

Model *Accuracy* *Standard Deviation*

```
LR: 0.749653 (0.004535)
LDA: 0.747307 (0.004110)
KNN: 0.717747 (0.004309)
CART: 0.789400 (0.005717)
NB: 0.738720 (0.003106)
```



CART eval:

	precision	recall	f1-score	support
0	0.58	0.59	0.58	6176
1	0.86	0.86	0.86	18824
accuracy			0.79	25000
macro avg	0.72	0.72	0.72	25000
weighted avg	0.79	0.79	0.79	25000

AUTHOR DETAILS

If you have any further questions, please contact the author at the following address:

Author: Clifton Lee
Email: Clifton.x.lee@gmail.com
LinkedIn: www.linkedin.com/in/clifton-lee/