Predicting user's demographics.

By RMM - September 2018

Final Capstone.

This capstone corresponds to the final part of the Thinkful Data Science Bootcamp. You can visit this notebook at my GitHub repository.

Outline

- ➔ Introduction
- ➔ Dataset information
- ➔ Dataset analysis & exploration
- ➔ Data transformation
- ➔ Training, testing and evaluation

Intro

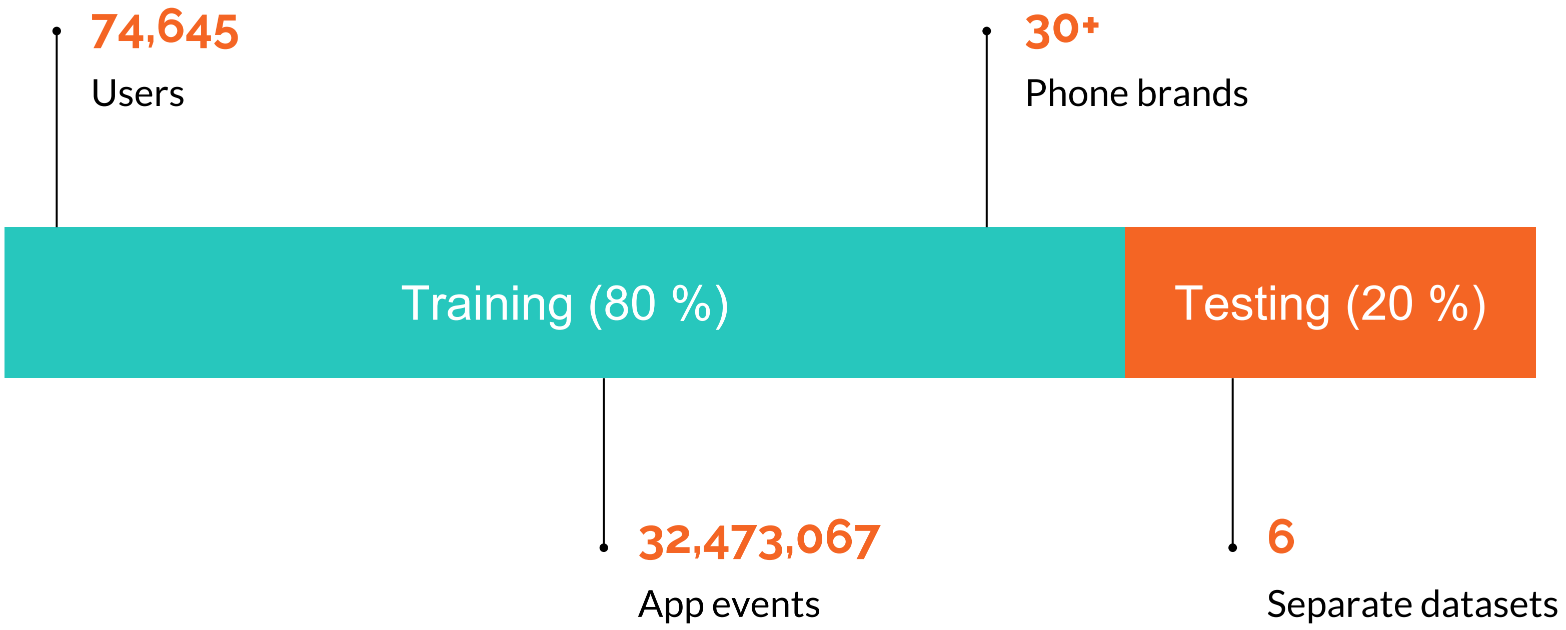
We'll explore and model a dataset from China's largest third-party mobile data platform, TalkingData.

Understanding that everyday choices and behaviors draw a picture of who we are and what we value, we can use this information to optimize apps and manage marketing resources.

The dataset contains information regarding app usage, geolocation and mobile device properties.

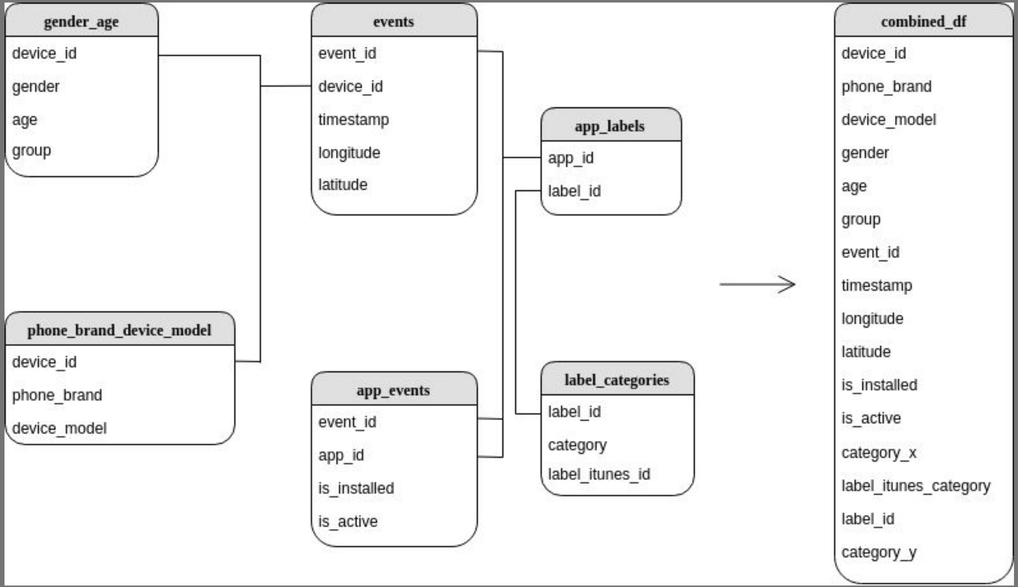
My objective is to predict user's demographic characteristics, gender, based on the latter information, which in return gives us valuable information that can help developers and brand advertisers around the world pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

Dataset information - TalkingData mobile data platform



Dataset Structure.

6 separate files build this dataset.

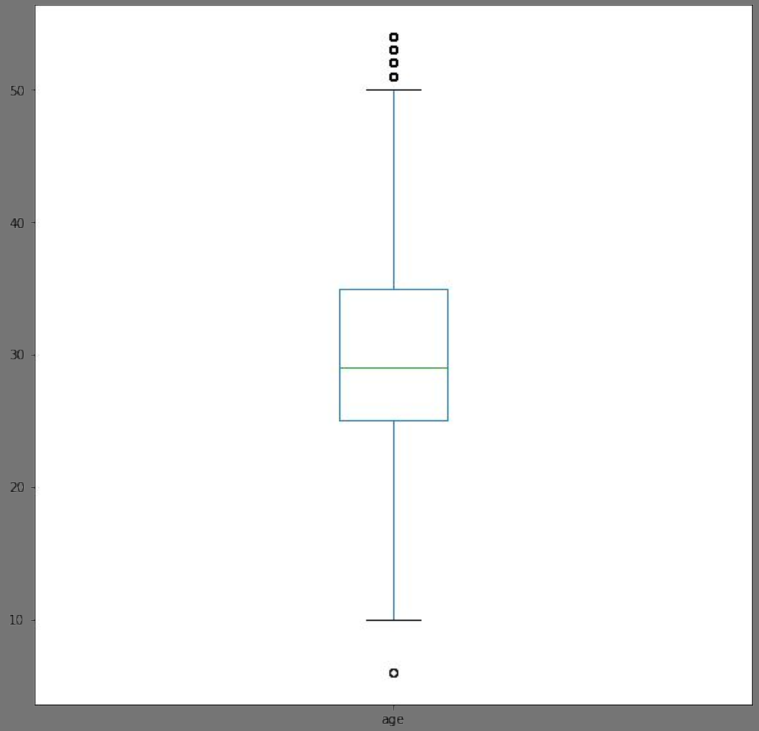


Due to the size of the combined dataset, a sampling function was used to reduce and split the original datasets.

Dataset
Exploration.

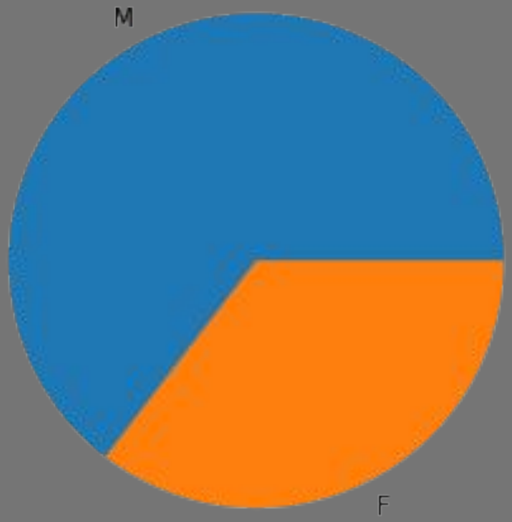
Age

72,486 users
Average age = 30
Concentrated
age between 25 & 35



Gender

72,486 users
Female users = 35 %
Male users = 65 %

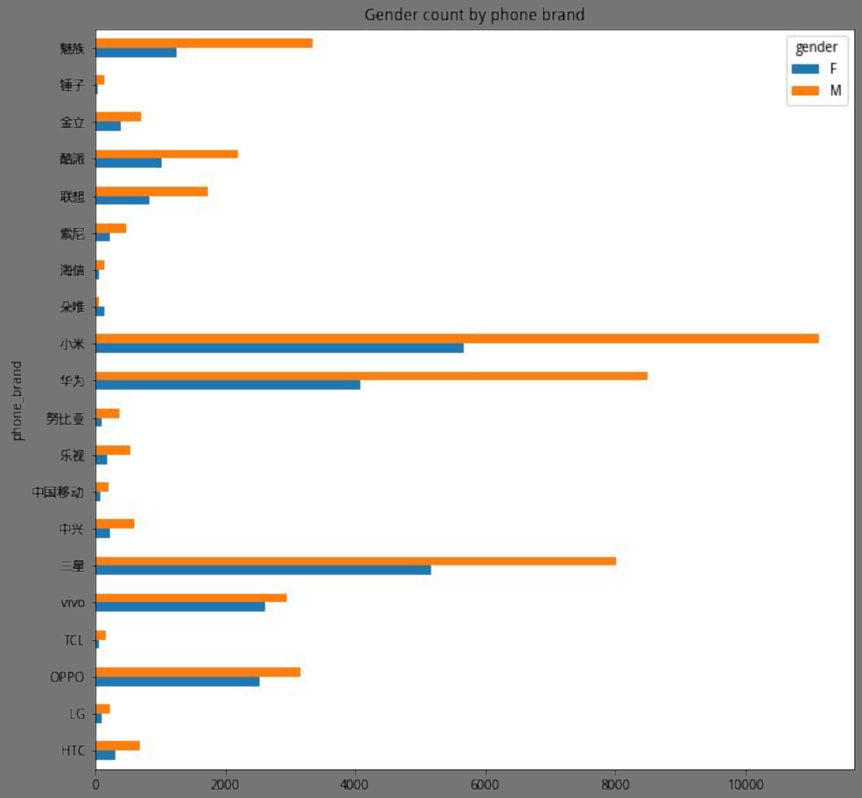


Top 3 phone brands

小米 (Xiaomi)
三星 (Samsung Group)
Huawei (华为).

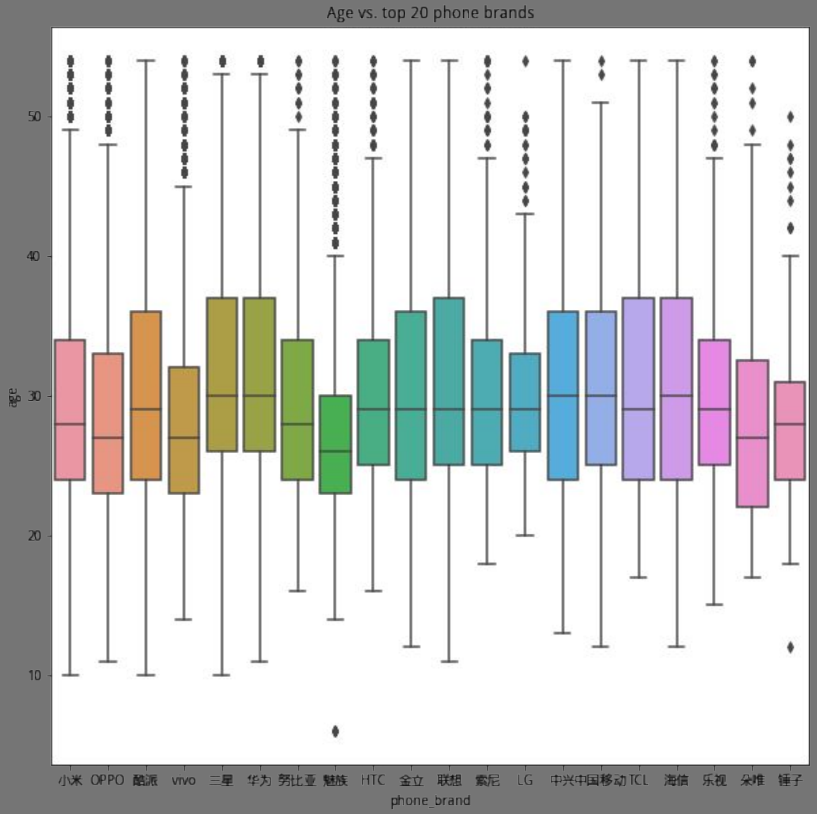
Gender distribution among top 20 phone brands

Male users tend to double the size in relation to female users



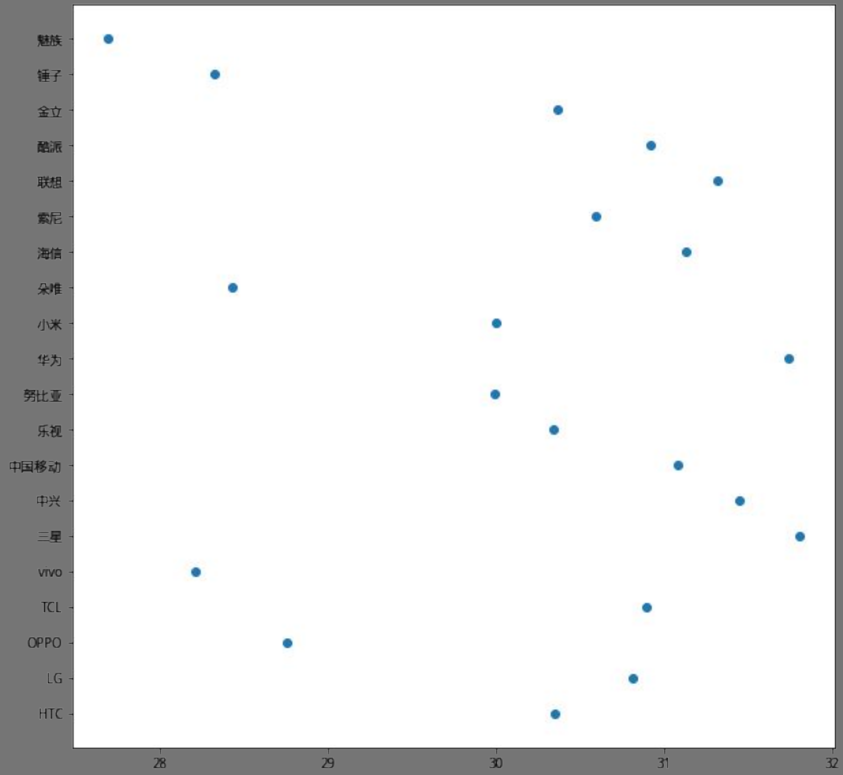
Age distribution among top 20 phone brands

Average tends to be 30 years



Age distribution among top 20 phone brands

Watching closer, 5 top brands have users below 29 years



Predicting Gender.

Dataset information - TalkingData mobile data platform

33,545

Users

17

Combined models tested

Training (80 %)

Testing (20 %)

85

Features

3

Techniques to enhance hyper parameters within each model

Creating dummy variables and balancing the dataset.

Creating dummy variables

Original features: 14

Current features: 85

Unbalanced training data

Synthetic Minority Over-sampling Technique (SMOTE)

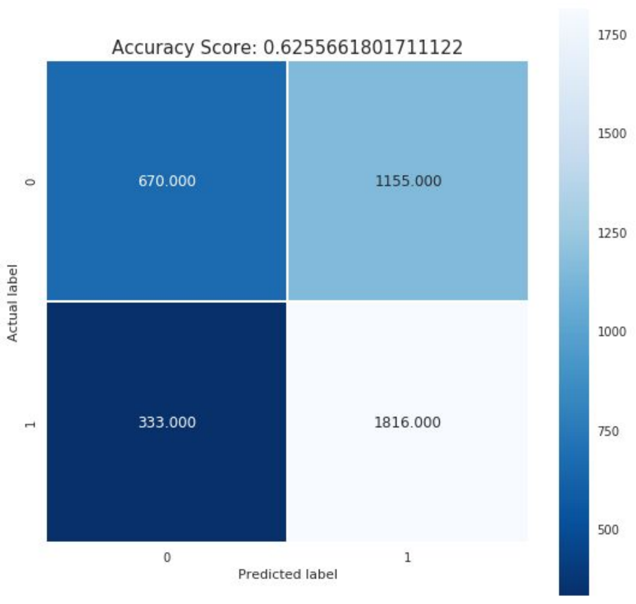
Balanced training data

Hyperparameter tuning: GridSearchCV

Logistic regression (Classification L2) classification report:

	precision	recall	f1-score
Male	0.67	0.37	0.47
Female	0.61	0.85	0.71
avg / total	0.64	0.63	0.60

Accuracy Scores - Test Set: 0.6256



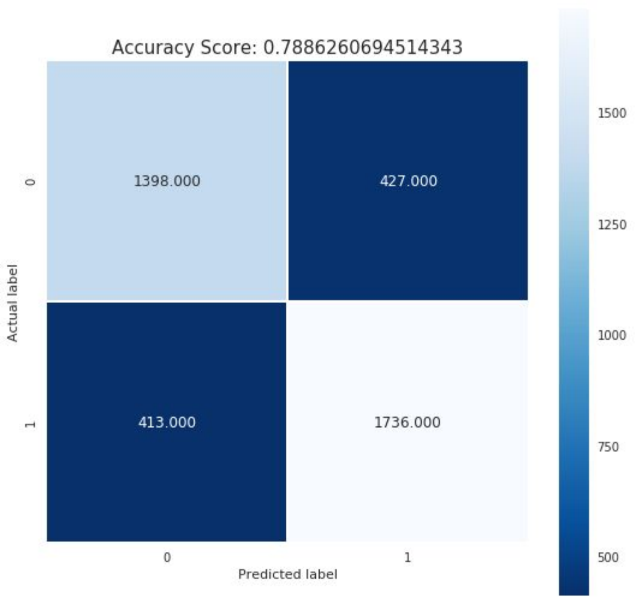
grid.best_score_ : 0.623026108839
grid.best_params_: {'C': 1}

Hyperparameter tuning: GridSearchCV

Random Forest Classification classification report:

	precision	recall	f1-score
Male	0.77	0.77	0.77
Female	0.80	0.81	0.81
avg / total	0.79	0.79	0.79

Accuracy Scores - Test Set: 0.7886



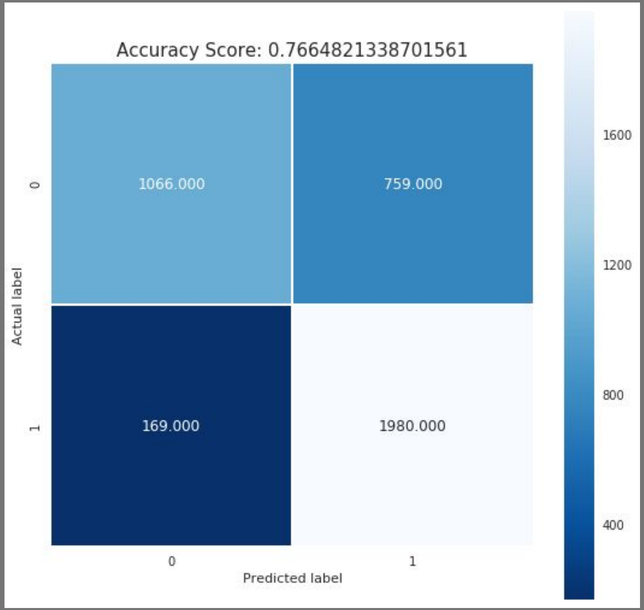
grid.best_score_ : 0.79811261403
grid.best_params_: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 1}

Hyperparameter tuning: GridSearchCV

Decision Tree Classifier classification report:

	precision	recall	f1-score
Male	0.86	0.58	0.70
Female	0.72	0.92	0.81
avg / total	0.79	0.77	0.76

Accuracy Scores - Test Set: 0.7664



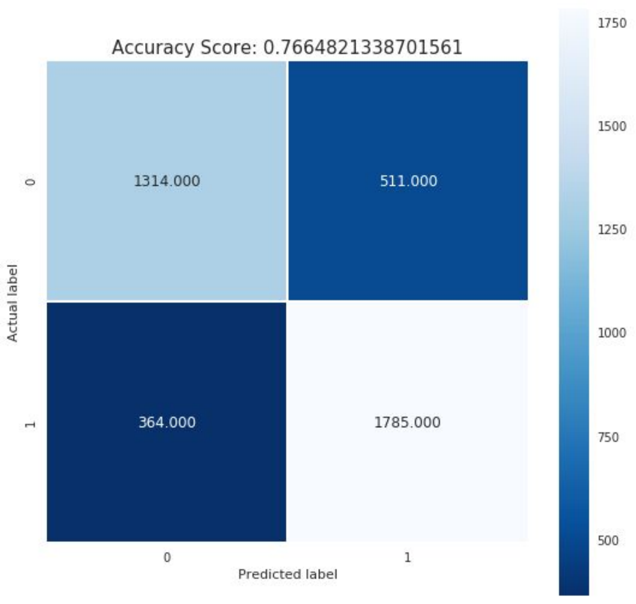
grid.best_score_ : 0.81981755269
grid.best_params_: {'criterion': 'gini', 'max_depth': 19, 'min_samples_split': 10}

Hyperparameter tuning: GridSearchCV

K-Nearest Neighbours Classifier classification report:

	precision	recall	f1-score
Male	0.78	0.72	0.75
Female	0.78	0.83	0.80
avg / total	0.78	0.78	0.78

Accuracy Scores - Test Set: 0.7664



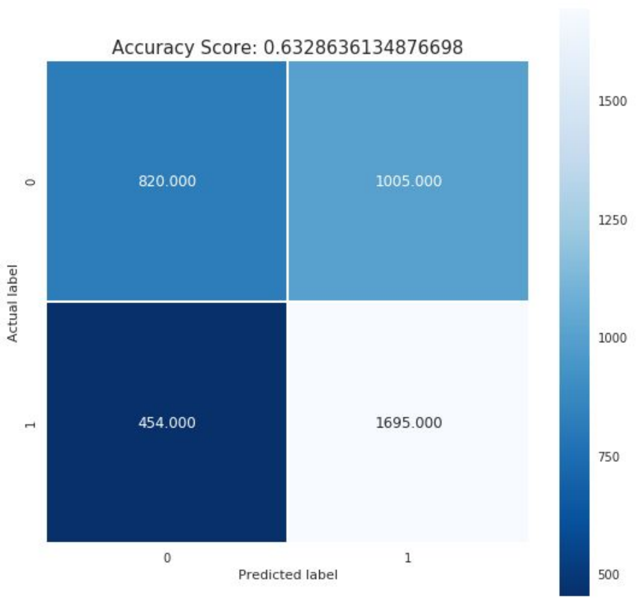
grid.best_score_ : 0.776659326832
grid.best_params_: {'n_neighbors': 7}

Hyperparameter tuning: GridSearchCV

Ridge Classifier classification report:

	precision	recall	f1-score
Male	0.64	0.45	0.53
Female	0.63	0.79	0.70
avg / total	0.64	0.63	0.62

Accuracy Scores - Test Set: 0.6329



grid.best_score_ : 0.623843976093
grid.best_params_: {'alpha': 10}

Predicting gender using PCA.

83 to 53

Fitting PCA to the training matrix, and retaining 75 % of it's variance we reduced the number of features used from 83 to 53 and optimizing prediction score and computational time.

Applying PCA

Logistic regression (Classification L2) classification report:

	precision	recall	f1-score
Male	0.55	0.09	0.16
Female	0.55	0.94	0.69
avg / total	0.55	0.55	0.45

Accuracy Scores - Test Set: 0.5488

Applying PCA

Random Forest Classification classification report:

	precision	recall	f1-score
Male	0.80	0.77	0.79
Female	0.81	0.84	0.83
avg / total	0.81	0.81	0.81

Accuracy Scores - Test Set: 0.8080

Applying PCA

Decision Tree Classification classification report:

	precision	recall	f1-score
Male	0.75	0.72	0.73
Female	0.77	0.79	0.78
avg / total	0.76	0.76	0.76

Accuracy Scores - Test Set: 0.7597

Applying PCA

K-Nearest Neighbours Classification classification report:

	precision	recall	f1-score
Male	0.80	0.72	0.76
Female	0.78	0.84	0.81
avg / total	0.79	0.79	0.79

Accuracy Scores - Test Set: 0.7881

Applying PCA

Ridge Classification classification report:

	precision	recall	f1-score
Male	0.55	0.09	0.16
Female	0.55	0.94	0.69
avg / total	0.55	0.55	0.45

Accuracy Scores - Test Set: 0.5488

Predicting gender using PCA & KMeans Clusters.

>>

PCA is used before running the K-means model, and finally predictions

Applying PCA & KMeans Clusters.

Logistic Regression (Classification L2) classification report:

	precision	recall	f1-score
Male	0.62	0.51	0.56
Female	0.64	0.73	0.68
avg / total	0.63	0.63	0.62

Accuracy Scores - Test Set: 0.6291

Applying PCA & KMeans Clusters.

Random Forest Classification classification report:

	precision	recall	f1-score
Male	0.77	0.78	0.77
Female	0.81	0.80	0.80
avg / total	0.79	0.79	0.79

Accuracy Scores - Test Set: 0.7906

Applying PCA & KMeans Clusters.

Decision Tree Classification classification report:

	precision	recall	f1-score
Male	0.77	0.80	0.79
Female	0.83	0.80	0.81
avg / total	0.80	0.80	0.80

Accuracy Scores - Test Set: 0.8004

Applying PCA & KMeans Clusters.

K-Nearest Neighbours Classification classification report:

	precision	recall	f1-score
Male	0.73	0.72	0.72
Female	0.76	0.77	0.77
avg / total	0.75	0.75	0.75

Accuracy Scores - Test Set: 0.7471

Applying PCA & KMeans Clusters.

Ridge Classification classification report:

	precision	recall	f1-score
Male	0.62	0.51	0.56
Female	0.64	0.73	0.68
avg / total	0.63	0.63	0.62

Accuracy Scores - Test Set: 0.6293

Predicting gender using Neural Nets.

>>

Sequential model

We used a sequential model and KerasClassifier to predict our user’s gender

Applying Neural Networks.

Individual dataframes were created to run the singular and the ensembled models respectively:

layers	5
model	Sequential()
epochs	3/5/7/8/9

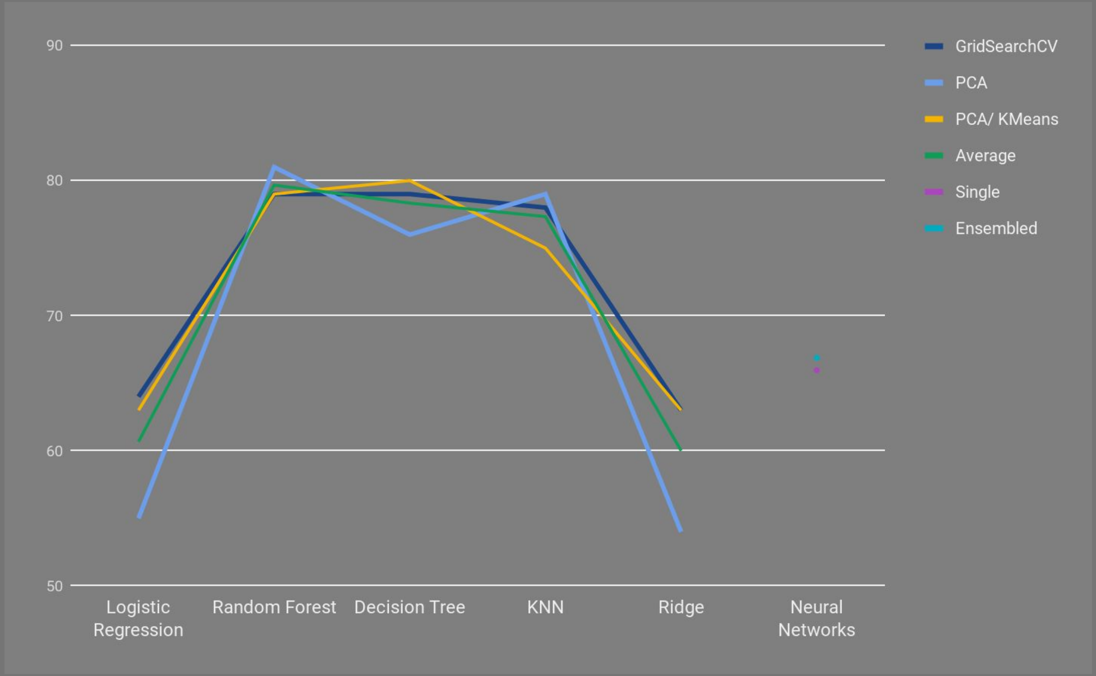
1-model Multilayer perceptron result:

Accuracy Scores - Test Set: 0.6597

5-model ensembled Multilayer perceptron result:

Accuracy Scores - Test Set: 0.6688

Overall Performance



Overall Performance

	GridSearchCV	PCA	PCA/KMeans	Single	Ensembled
Logistic Regression	64.0	55.0	63.0		
Random Forest	79.0	81.0	79.0		
Decision Tree	79.0	76.0	80.0		
KNN	78.0	79.0	75.0		
Ridge C.	63.0	54.0	63.0		
Neural Networks				65.97	66.88

Best model performance for predicting gender.

>>

Random Forest C.	81.00 %
Decision Tree	80.00 %
KNN	79.00 %
Neural Networks	66.88 %
Logistic Regression	64.00 %
Ridge C.	63.00 %

How can this information help us?

Marketing strategies

Advertising campaigns are user-oriented

Knowing your user

App developers can shape their software accordingly and generate higher ROI

R&D investments maximized

Investments go were matters and savings are wisely executed

TalkgingData mobile data platform

Aspects we could improve to increase our predictability.

>>

- Increase power
- Hyper parameter tuning
- Multiple algo testing
- Ensemble + models
- Multi target classification

Thank you Hoa Tran.
Thank you.