EU restaurants classifier

Module 5 project
N. Kairakbayev

Trip Advisor EU Cities Restaurants Dataset

Dataset size

- **Samples: 125K**
  - **78K – after cleaning**
- **Features: 10 features**
  - **160 – after parsing**

Features after parsing

- **Ranking ~ 16K unique**
- **Rating ~ 11 unique**
- **Number of reviews ~ 2K unique**
- **Price Range (target) ~ 3 unique**
- **City ~ 31 unique**
- **Cuisine ~ 125 types with different combs for each sample**

Source: https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw

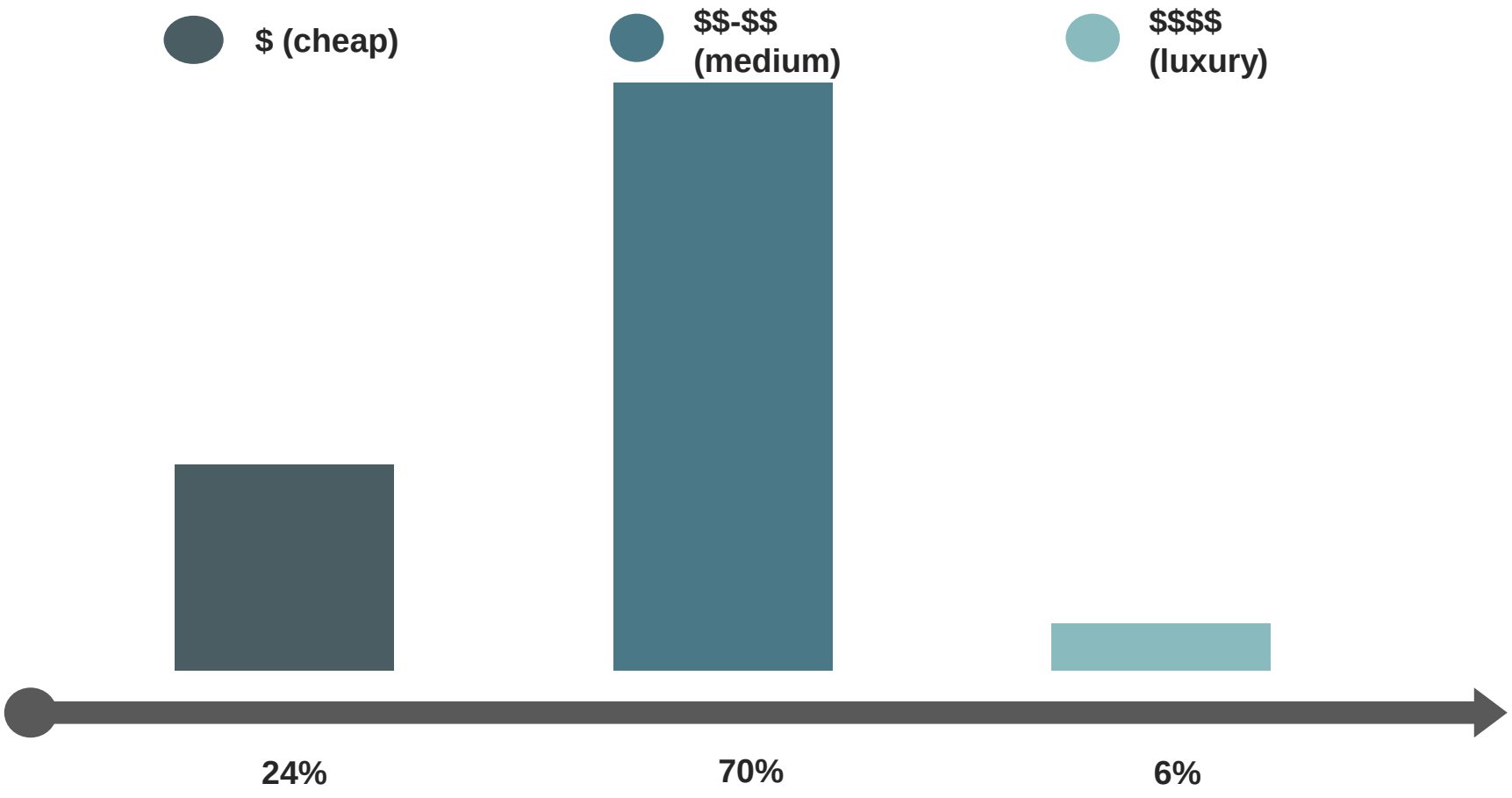
Key Question

What are the factors defining price range?

Key aspects

- **Features**
- **Validation metrics**
- **Classification algorithm**
- **Benefits for business**

Price Ranges

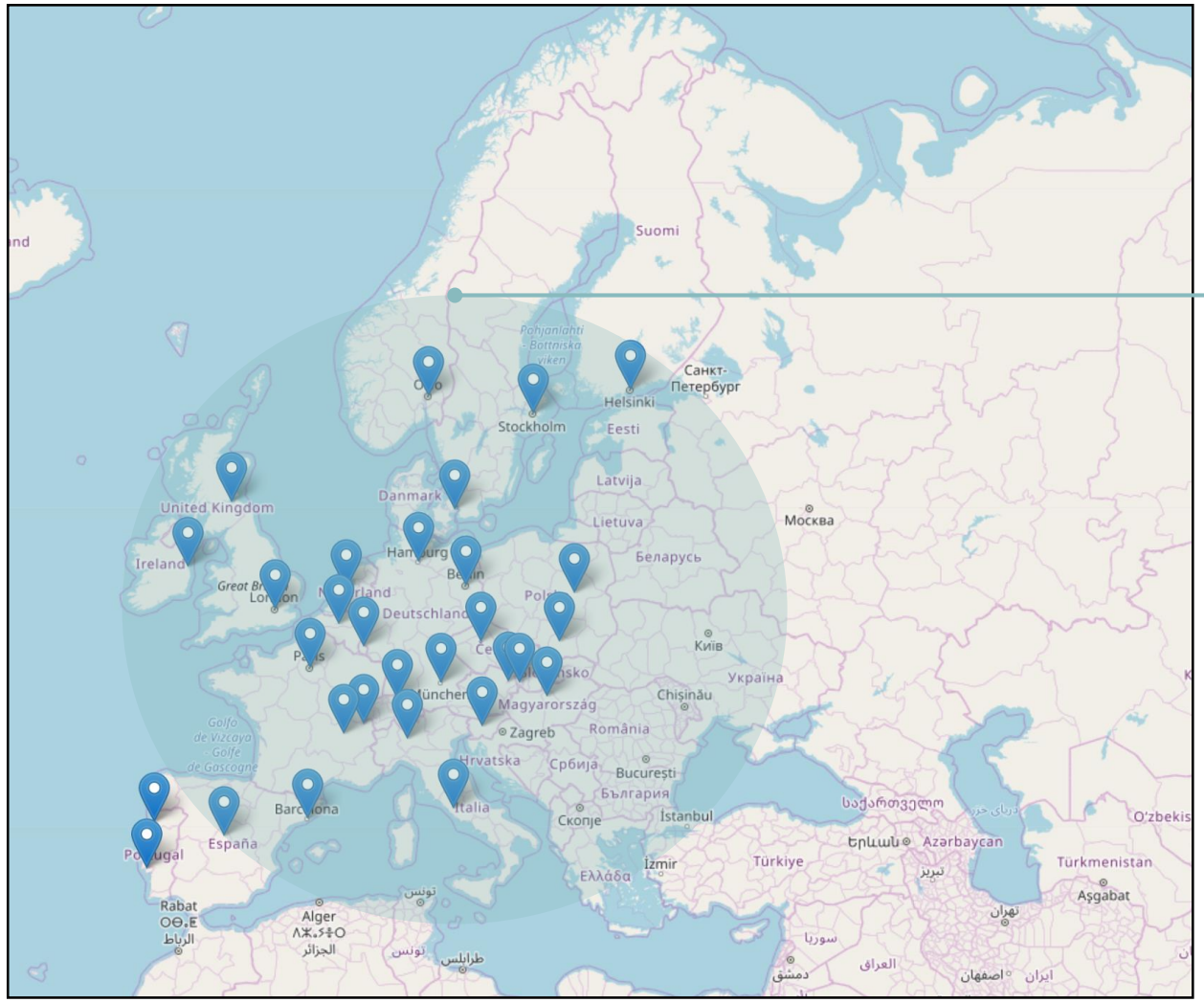


19 005	54 360	4 307

TOP3 restaurant statistics

	CHEAP ($)	MEDIUM ($$-$$$)	LUXURY ($$$$)
TOP3 Cities	1. London 2. Paris 3. Roma	1. London 2. Paris 3. Roma	1. London 2. Paris 3. Barcelona
TOP3 Cuisine Styles	1. Vegetarian Friendly 2. European 3. Italian	1. Vegetarian Friendly 2. European 3. Mediterranean	1. European 2. Vegetarian Friendly 3. Gluten Free
Average Rating	4.1	4.0	4.2

Cities Map



- **31 cities**
- **23 capitals**
- **24 countries**

Evaluation Process

- **Main Metric: Accuracy Score**

- **Data Split:**
  - Kfold validation: for Grid Search
  - Train-test-split: for Final Model selection

DATA CLEANING and EDA

- Missing data
- Parsing cuisine and binning
- Visual analysis
- Correlation matrices

PREPROCESSING

- PCA
- Clustering

GRID SEARCH

only on TRAIN SET with KFOLDS validation

FINAL CLASSIFIER MODEL

Final fit on TRAIN SET and validate on TEST SET

- Models (pipes) comparison
- Fit Chosen Classifier
- Monte-Carlo Simulation on accuracy

Applied Models and Methods

A. PCA

Correlation Matrix
Break-even VAR portion – 97,5%

B. CLUSTERING

1. Visual Analysis: 2D KDE PLOT
2. Cluster models:
   - Kmeans++
   - DBSCAN
3. Metrics:
   - Calinski-Harabaz
   - Silhouette score

C. Single Models

1. KNN
2. Decision Tree
3. SVM-Classifier
4. Logistic Regression
5. Naïve Bayes Classifier

D. Ensemble Models

1. Random Forest
2. ADA-Boost
3. XG-Boost

E. Ranking metrics

1. Accuracy Score
2. F1-score

E. Validation

1. Kfold – on Grid Search
2. Train-Test-Split – on Ranking
3. Pipelines – for process incapsulation
4. Monte-Carlo – for verify final score

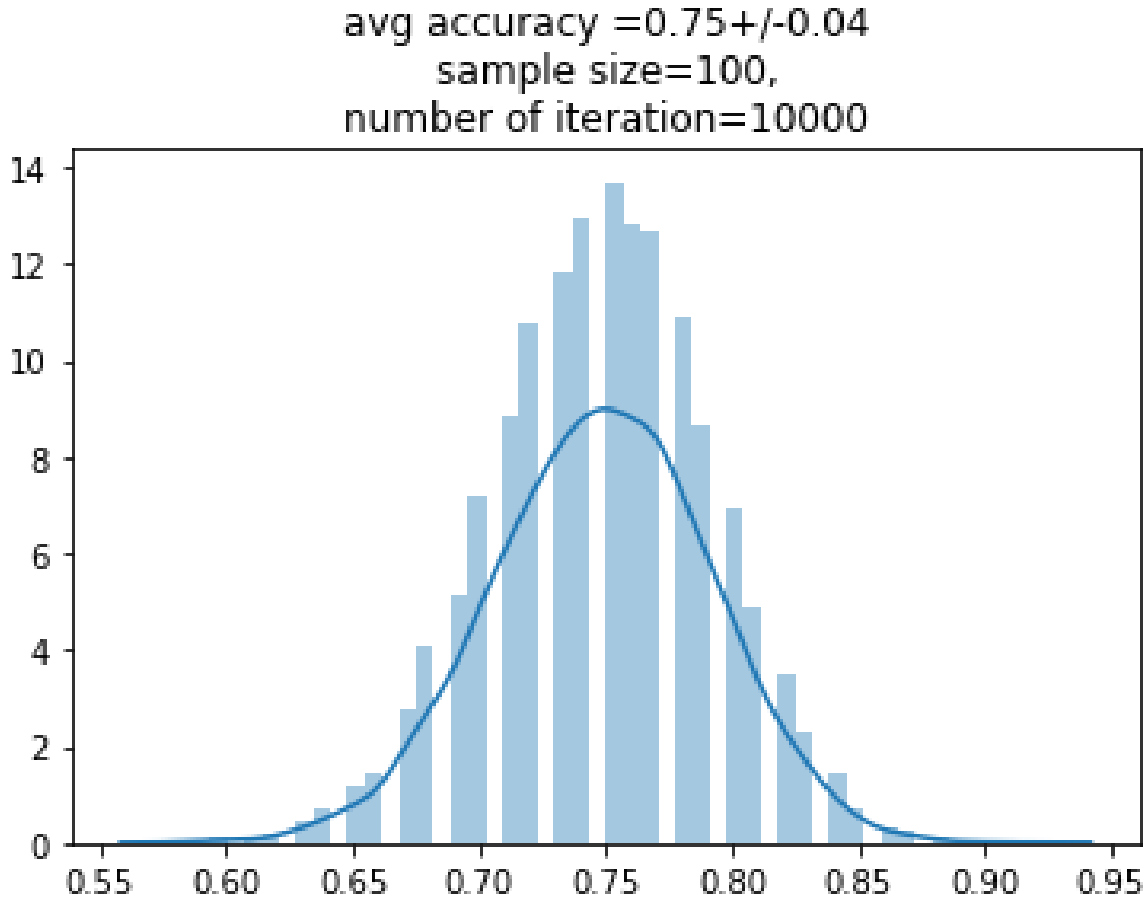
Models Comparison on Test set

SINGLE MODELS		ENSEMBLE MODELS	
Model	Accuracy Score	Model	Accuracy Score
KNN	0.737	Random Forest	0.738
Decision Tree	0.727	ADA-Boost	0.736
SVM	0.739	XG-Boost	0.749
Logistic Regression	0.742		
Naïve Bayes	0.696		

Best model: XG-Boost:

- **0.749 accuracy score on test set**
- **0.800 accuracy score on train set**

Final Classifier



Business Recommendations

- Restaurant **Price Ranges** could be **explained** and **predicted** by its Citiy, Rating, Ranking and Cuisine Styles with moderate level of accuracy (**75%**)

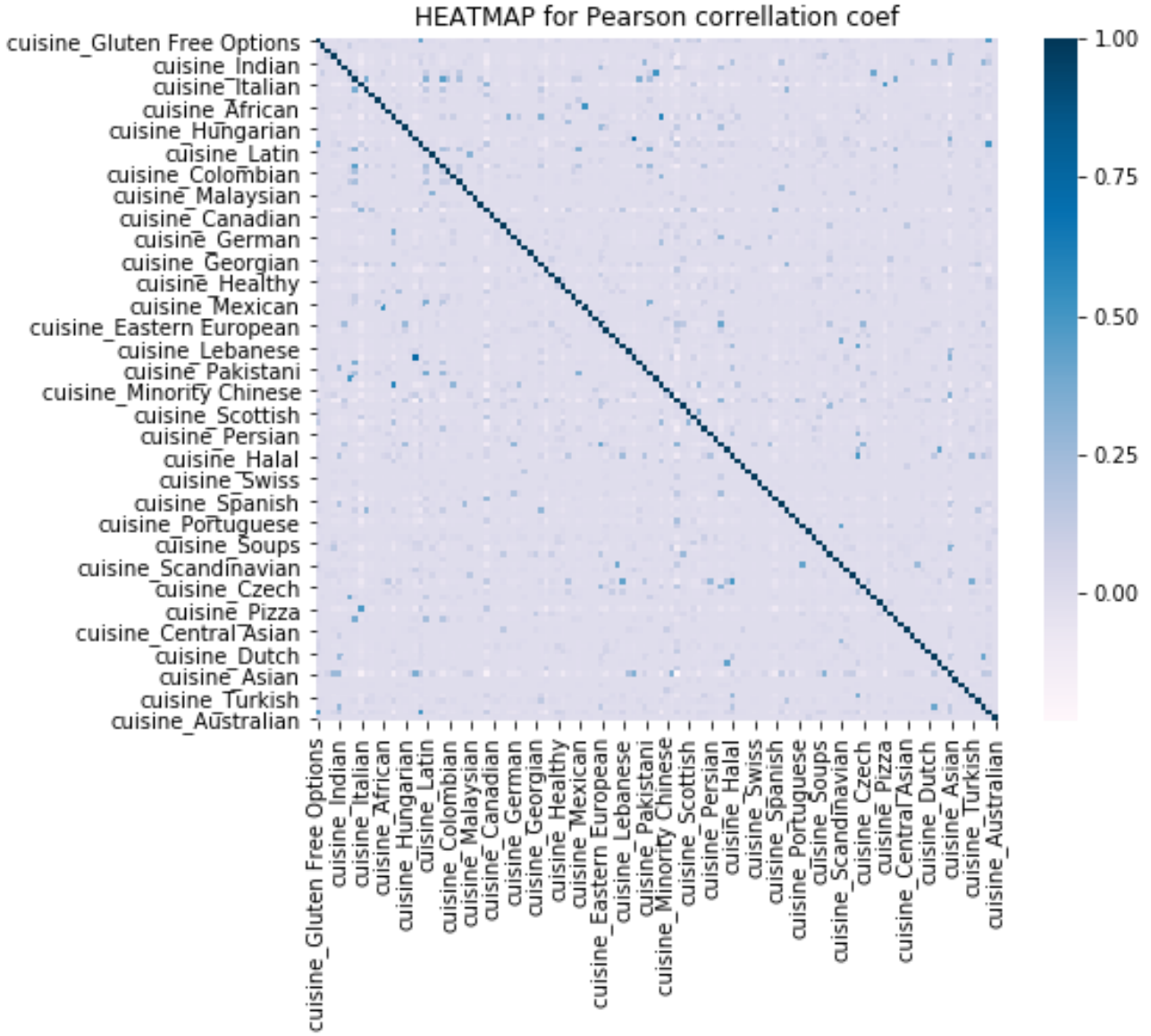
- **Classifier Model** provides to do **Market Segmentation and Research** on EU Restaurant business sector

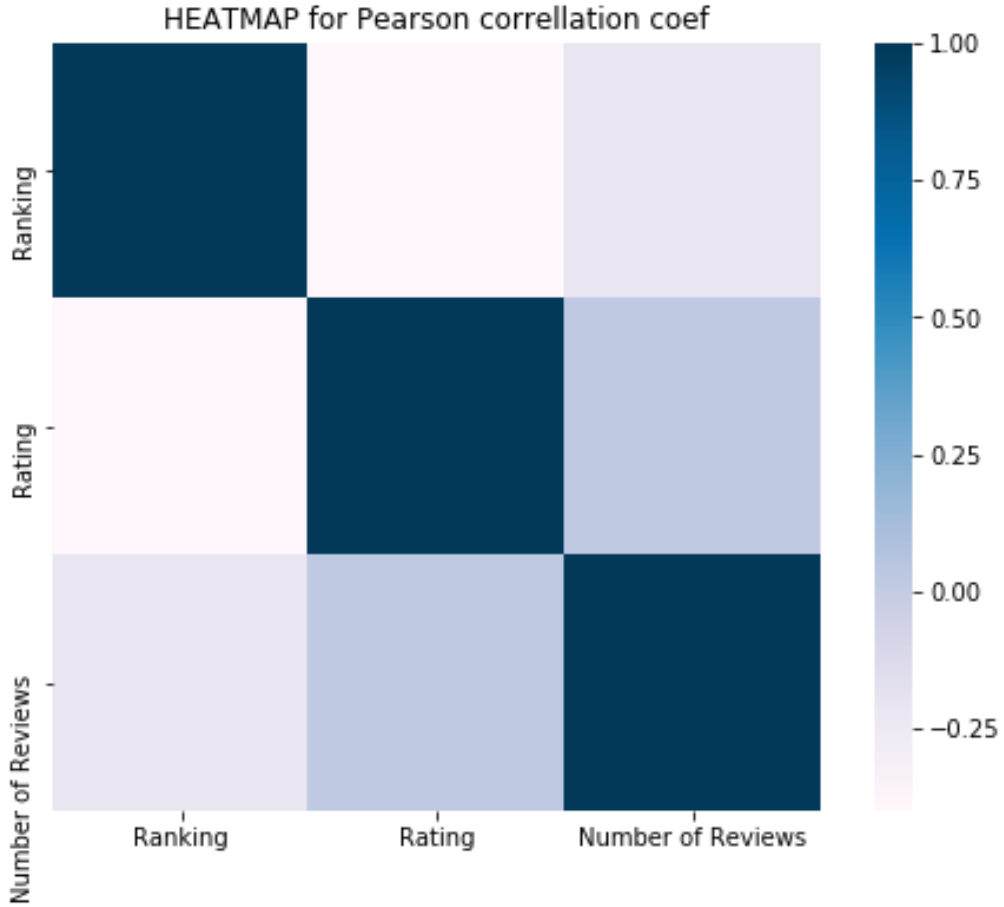
- Classifier Model provides **scenario inputs for financial evaluation** of acquisitions and new restaurant project

Future Work

1. **Find new predictors and samples on missing data and update Classifier**
2. **Add NPL analysis on Reviews**
3. **Apply Deep Learning methods**

Appendix A: Correlation matrices

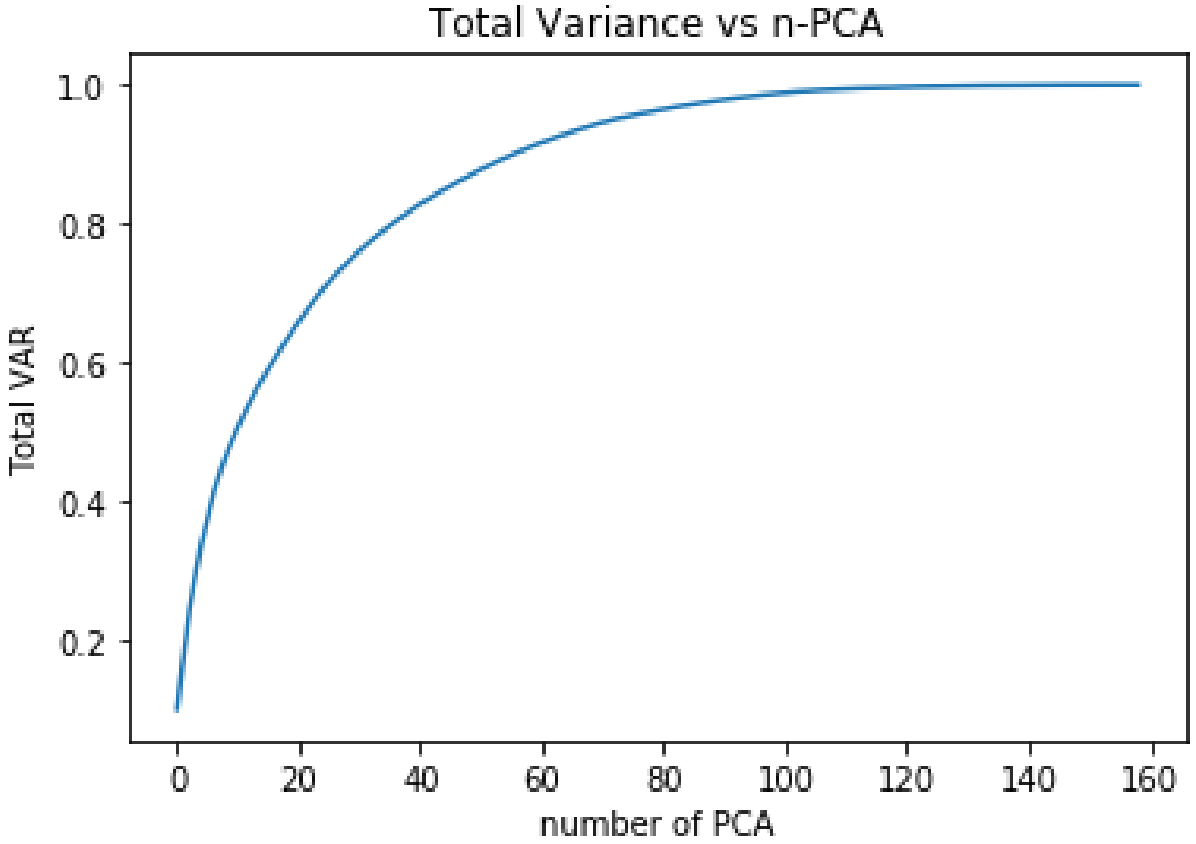




Appendix B: PCA

% of Total VAR	# of PCA
99%	102
97.5%	87
95%	72
90%	56
80%	36

Initial X size: 159



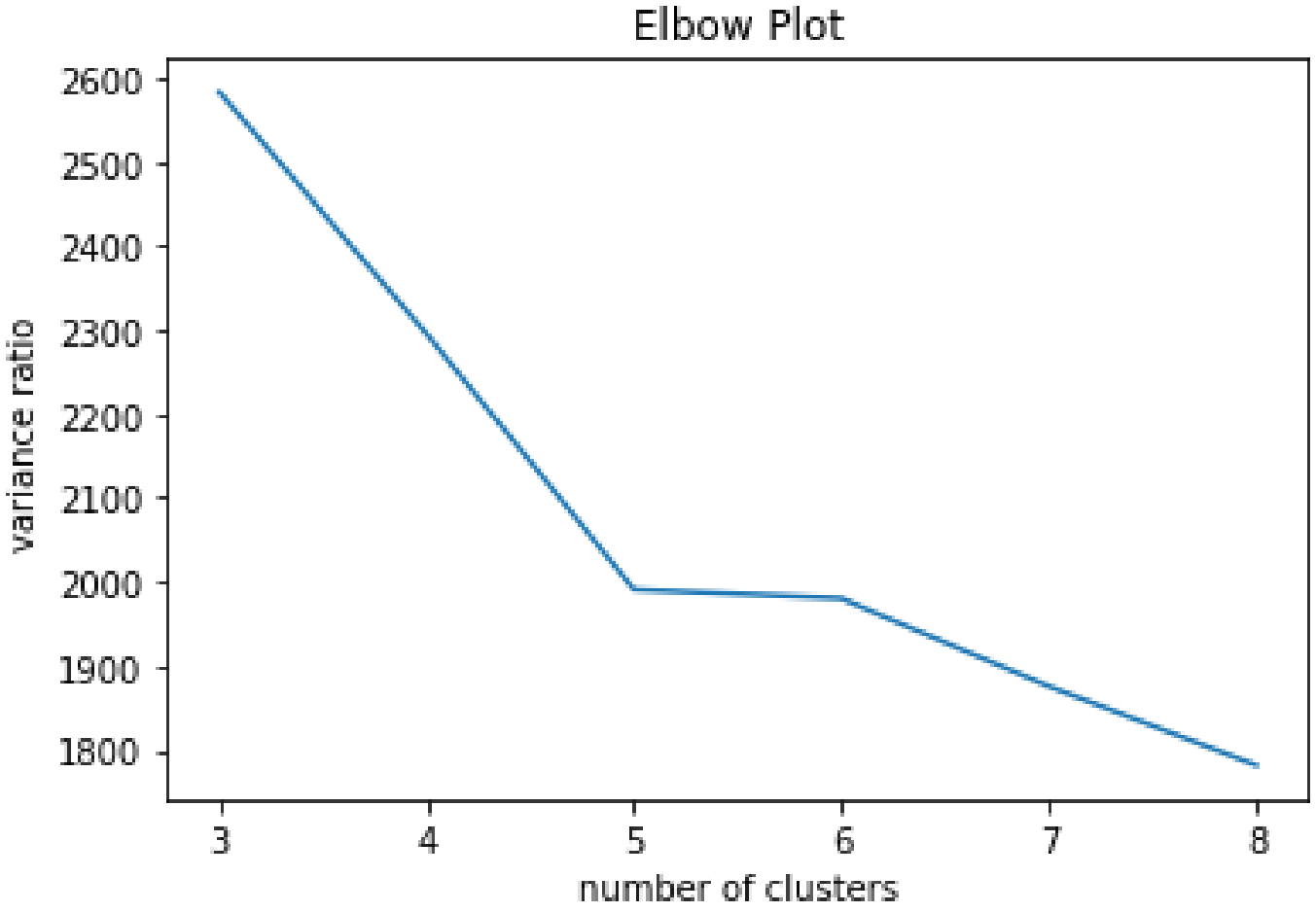
Appendix C: Clustering Analysis

Kmeans++

Silhouette Score – 0. 06 on N = 3

DBSCAN

Silhouette Score: varies from -0.25 to 0.19



Appendix D: params for single models

KNN

- n_neighbors: 50
- Algorithm: auto

CART

- Criterion: gini
- Max_depth: 10
- Max leaf nodes: 50
- Min sample split: 80

SVM

- SVM: LinearSVM
- C: 1
- Max_iter: 100
- tol: 1e-12
- penalty: l2
- Multi_class: ovr

LogReg

- Solver: SAGA
- C: 1
- tol: 1e-8
- penalty: l2
- Max_iter: 100

Appendix D: params for ensemble models

Random Forest

- Criterion: gini
- n_estimator: 15
- max_depth: 10
- Min_sample_split: 2

ADA-boosting

- Algorithm: SAMME.R
- Learning_rate: 0.3
- n_estimator: 1000

XG-boosting

- Booster: gbtree
- n_estimator: 100
- learning_rate: 0.1
- Max_depth: 7
- Base_score: 0.75
- reg_lambda: 1