Analysis of Brand Image on Instagram

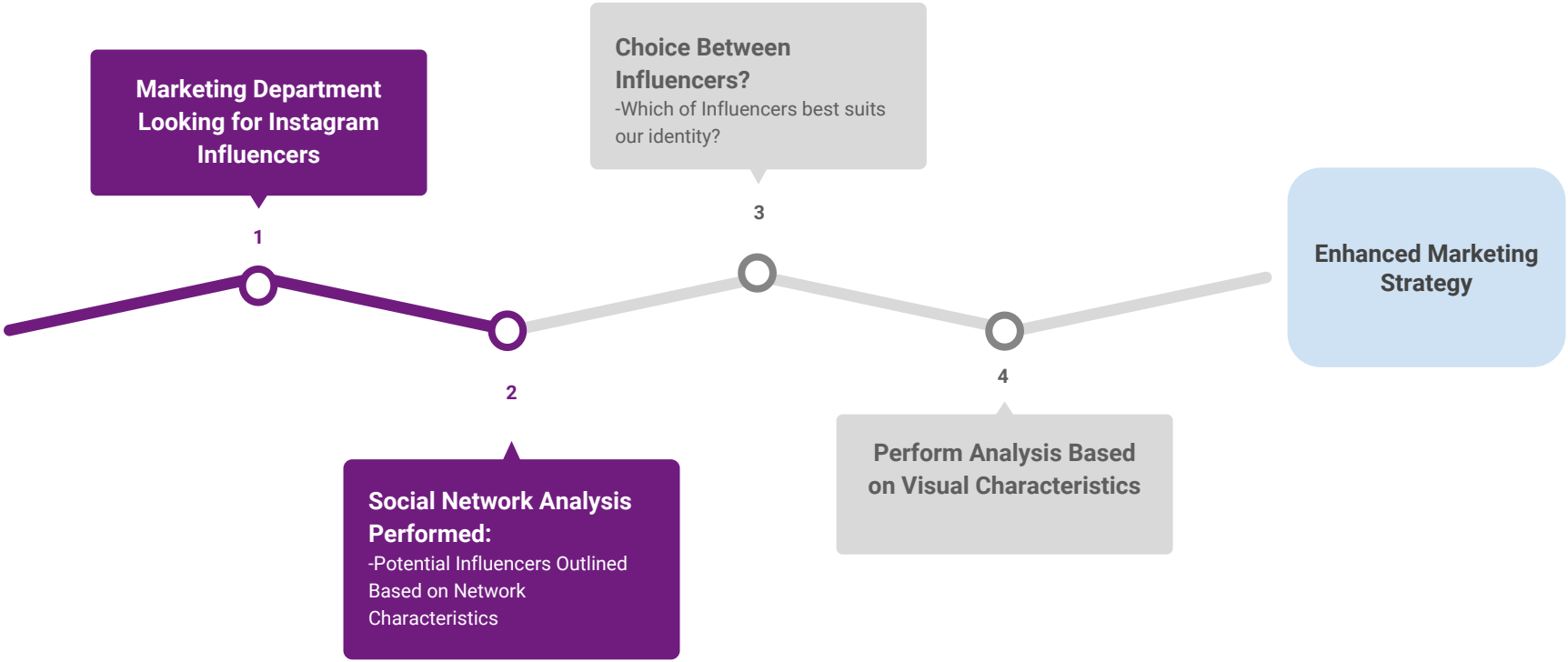
Omar ALShaye, Matthew Rosenthal, Lisa Lee,
Ann Eitrheim, Matthew Echols

Business Problem and Goals

- Key question and Goal:
  - What instagram images identify most closely with a brand?
  - Predict brand category based on visual features of instagram posts

- Insights:
  - Develop an instagram advertising and targeting strategy focusing on synergies between brand groups and influencers to complement network analysis

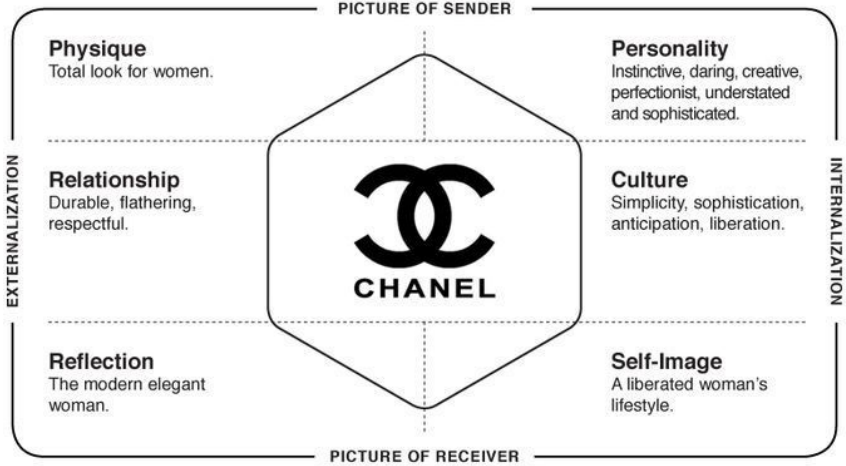
Business Value: Building a Complete Marketing Strategy



Kapferer Brand Identity Prism and Hypothesis

- Tool that connects/aligns **brand identity** (how brands want to be perceived) with **brand image** (how customers actually view the brand)

- Data analysis will be used to reinforce a brand's identity prism for more effective marketing



- **Our Hypothesis:** Instagram posts of the same brand category are expected to have similar aesthetics which can be leveraged for marketing
  - Customers wearing similar brands will form similar brand image
  - Instagram image features are closely tied to brand identity and can be used complement traditional marketing efforts targeting a brands core audience

Data Sources and Exploratory Statistics

Data Source

- Data was from Harvard Dataverse
- Fashion posts were accessed through the Instagram's API over the period of June 2015 - January 2016
- Posts mentioning particular hashtags such as fashion keywords and specific brand names. Posts were limited to a set containing at least one hashtag of a brand
- The brand list contained 48 internationally renowned names from luxury fashion houses (Hermes, Prada) to high street brands (Zara,Forever 21)

Key Features & Dimensions

Category	Features
Basic information about posting user and each post	User Id, Followings ,Followers, Media count, Brand name, Brand category, Hashtags, Caption, Image URL, Likes, Comments, Creation Time, Link
The visual content variables and descriptions	Selfie, Body snap, Marketing, Product-only, Non-fashion, Face, Logo, Brand logo, Smile, Outdoor, People, Items, Various emotions

Dimension: 24,752 rows (Posts) and 33 columns for 13,350 people on Instagram

Brand Categories

Mega Couture (3)

GUCCI

D&G
DOLCE & GABBANA

High Street (0)

ZARA

TOPSHOP

Small Couture (2)

Brioni

Kiton

Designer (1)

kate spade
NEW YORK

COACH

Visual Features

Image 1 (Selfie)

Image 2 (Body snap)

Image 3 (Marketing)

Image 4 (Product)

Image 5 (Non-fashion)

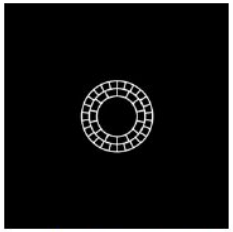


Image 1 (Others)

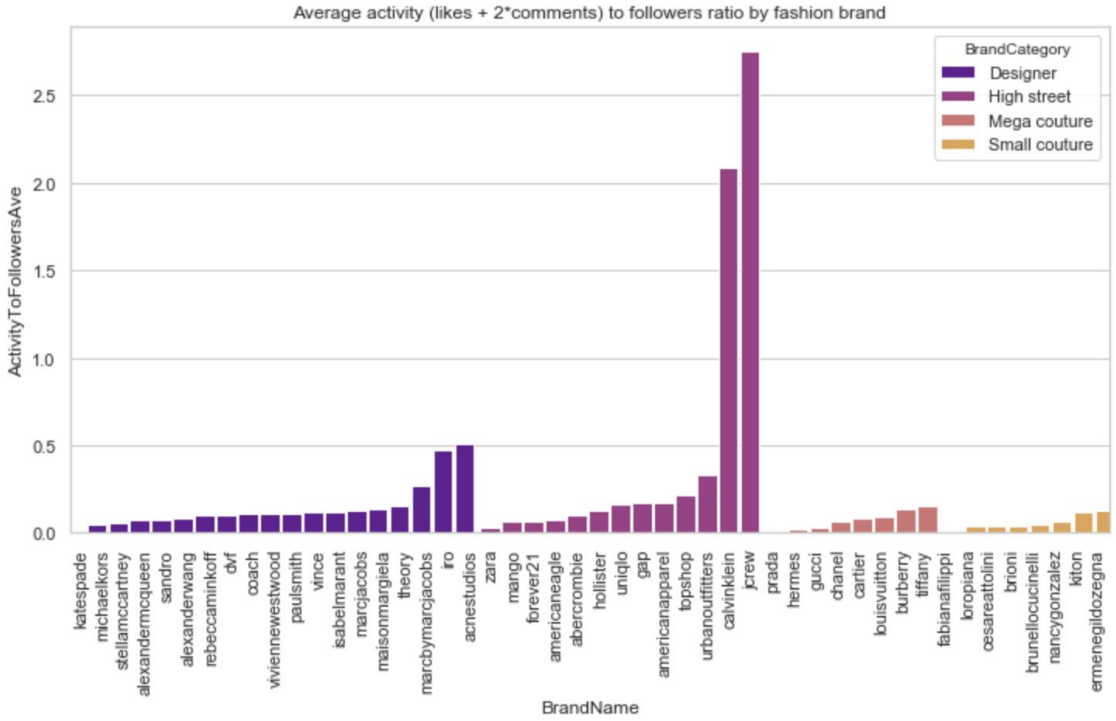
Image 2 (Advertisement)

Image 3 (Textile)

Image 4 (Marketing)

Image 5 (Multifunctional)

Exploratory Statistics



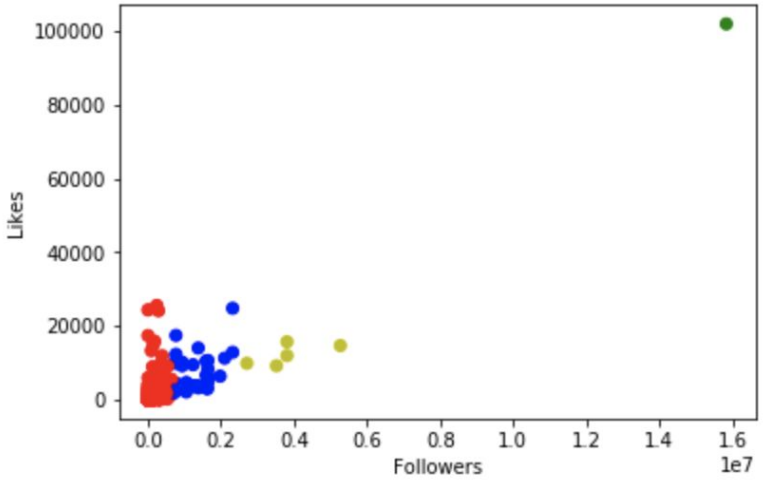
- This variable demonstrates a brands effectiveness as generating brand interactions with customers
- JCrew and Calvin Klein followers interact with posts more often than those of other brands and can serve as models for other brands

Modeling and Analysis

Unsupervised Methods: Clusters based on Brand Category

Attempted KMeans to cluster into groups of Brand category

- The results indicate no clear clustering for brand categories
- It seems there are only 2 clusters

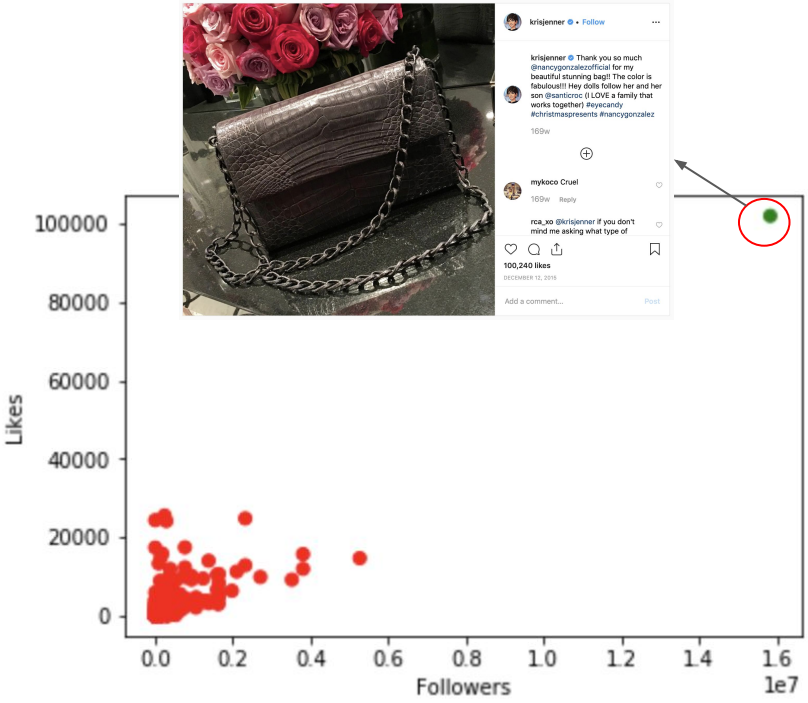


BrandCategory labels	Designer	High street	Mega couture	Small couture
0	11061	5770	1857	6025
1	0	0	0	1
2	5	0	0	0
3	20	3	0	10

Unsupervised Methods: Clusters based on Likes & Followers

- Outlier of high follow/like status: People who need not fit a specific brand image to attain traction

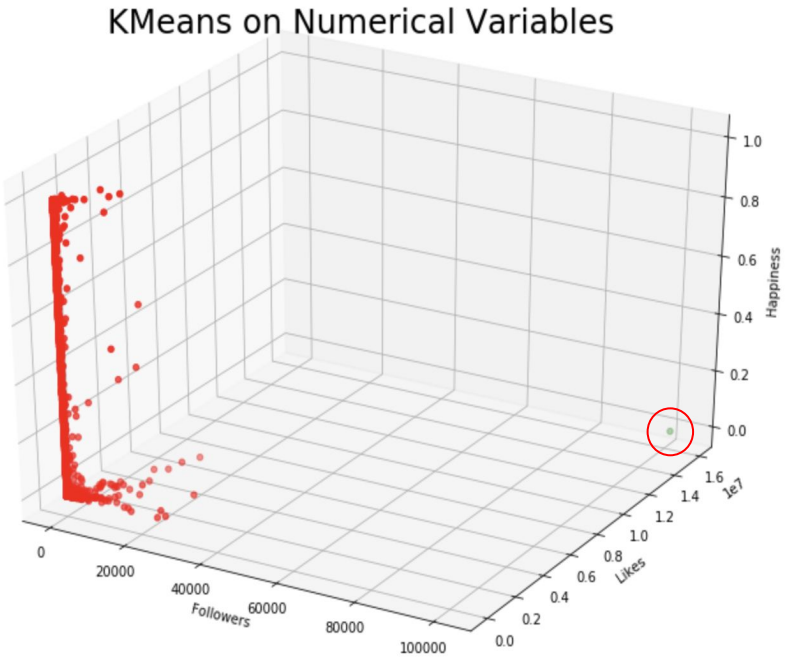
- Analysis on visual features will help choose among non-outlier potential influencers



Unsupervised Methods: Intuition on effect of Visual Features

A visual feature model will help companies choose between influencers who have lower levels of likes & followers

- The addition of visual features had greater impact among influencers with lower number of followers
- Popular/celebrity influencers received large numbers of likes regardless of visual features



Supervised Methods Overview

- Train/Test split of 25%
  - Train: 18,564 observations
  - Test: 6,188 observations
- Approach:
  - Feature Engineering
  - Assigned baseline model:
    - KNN, Decision Tree, SVC, Random Forest
  - Performed hyperparameter tuning to select the best parameters that maximize model accuracy

Feature Engineering

- Map Brand Name and Category to Integer Values
- Create Time Related variables from TimeStamp
  - Day of the Week
  - Hour
  - Season
  - WeekMod
- Time related features have higher feature importance:
  - Predicting Brand is highly dependent on seasonality and hour of the day

Feature Engineering (Continued)

- Non-time related Features
  - A binary variable for Smile and Negative Emotion Features (Eg. Disgust, Anger)
    - Bucketed values for negative emotion features
- Feature Selection
  - We left out Brand Name because it will cause our model to be trivial since brand name is directly associated with brand category
    - For eg, abercrombie's brand category is known to be high street fashion

Final Variables Selected

Y	X Features
Brand Category	'Selfie', 'BodySnap', 'Marketing', 'ProductOnly', 'NonFashion', 'Face', 'Logo', 'BrandLogo', 'Smile', 'Outdoor', 'NumberOfPeople', 'NumberOfFashionProduct', 'Anger', 'Contempt', 'Disgust', 'Fear', 'Happiness', 'Neutral', 'Sadness', 'Surprise', 'DayOfWeek', 'Hour', 'WeekMod', 'Season'

Baseline Model Comparison

Random Forest

```
random_forest = RandomForestClassifier(n_estimators=500, min_samples_leaf = 3, max_features = .5, n_jobs = -1)
random_forest.fit(train_X, train_Y)
random_forest_predictions = random_forest.predict(test_X)
```

Support Vector Machine Classifier

```
svc=SVC()
svc.fit(train_X, train_Y)
Y_pred2 = svc.predict(test_X)
```

K-Nearest Neighbors

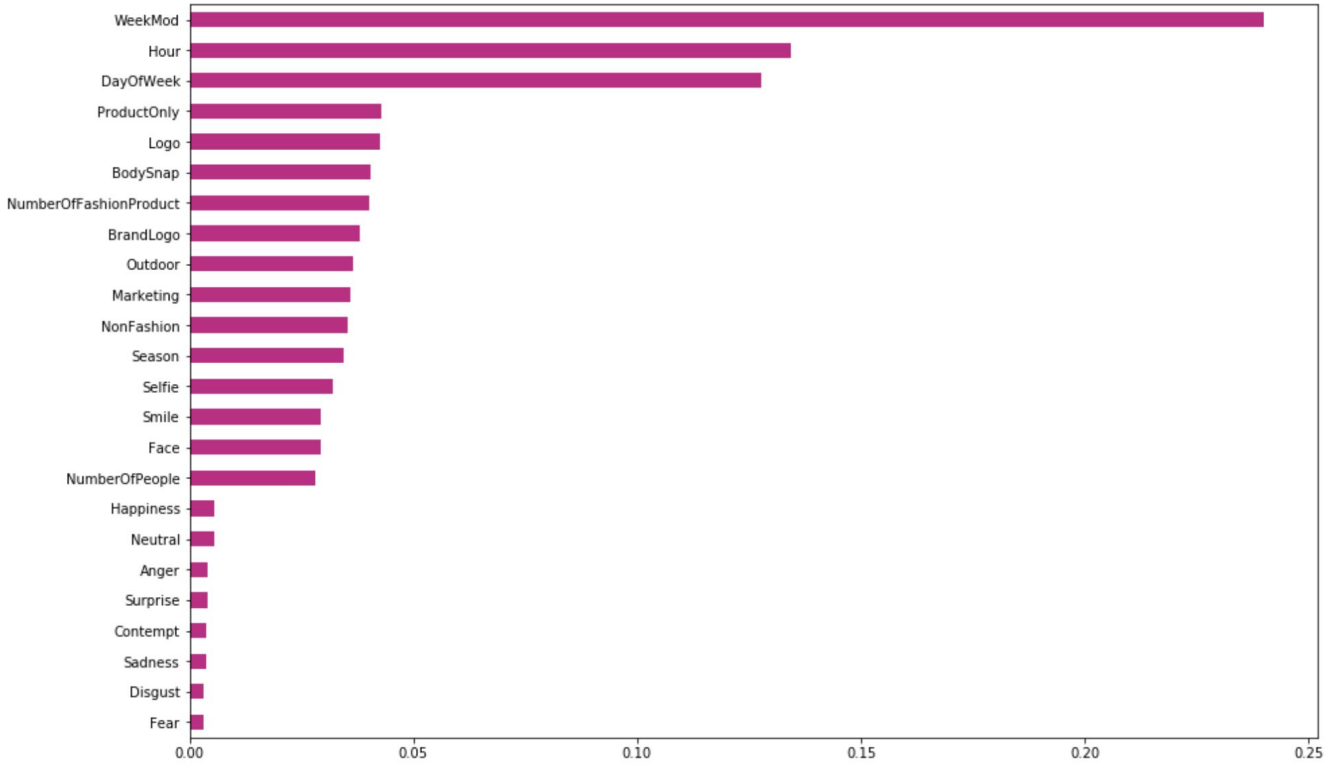
```
knn = KNeighborsClassifier(algorithm='auto', leaf_size=26, metric='minkowski',
                           metric_params=None, n_jobs=1, n_neighbors=10, p=2,
                           weights='uniform')
knn.fit(train_X, train_Y)
knn_predictions = knn.predict(test_X)
```

Performed Cross-Validation, with k = 10

- cross_val_score(MODEL, X, Y, scoring='accuracy', cv=10)

Model	Score
Random Forest	0.704570
Support Vector Machines	0.686292
KNN	0.618176

Feature Importance



Hyperparameter Tuning

Hyperparameter Tuning: Randomized Search with CV

- Trained randomized search on random forest model using baseline parameters
- Fit randomized search to data
- Identify best parameters for random forest based on trained randomized search
- Evaluate randomized search using best parameters and cross validated

```
{'n_estimators': 1400,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 80,
 'bootstrap': False}
```

Model Accuracy Score:

77.0

Hyperparameter Tuning: Grid Search with CV

- Create the parameter grid based on the results of randomized search
- Fit the grid search to the data
- Identify best parameters for random forest based on grid search
- Evaluate grid search using best parameters and cross validated

```
{'bootstrap': False,
 'max_depth': 100,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 1300}
```

Model Accuracy Score:

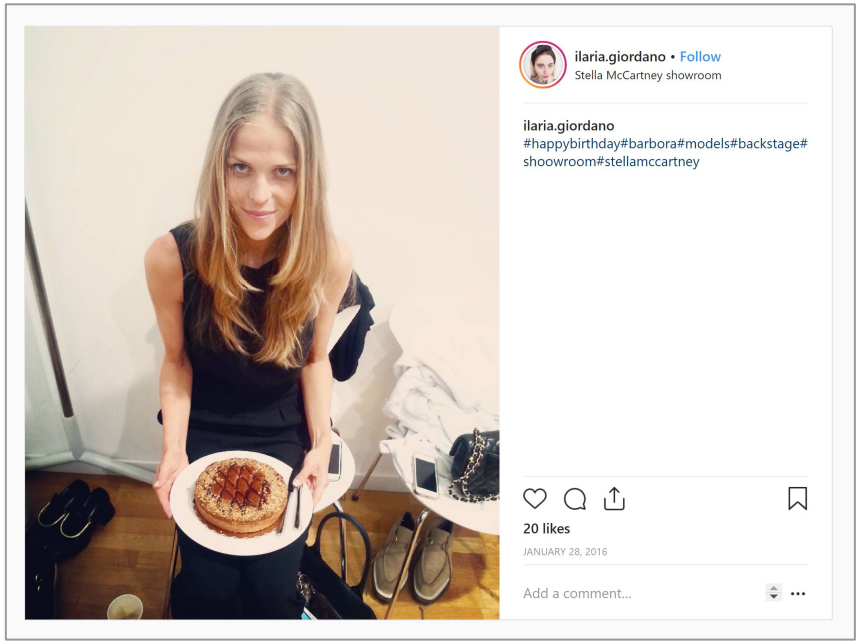
76.78

Insights

Real Life Validation:

UserId:
11723096675366714
50_1300829579

Designer:
Stella McCartney



Predicted:
1 (Designer)

Actual:
1 (Designer)

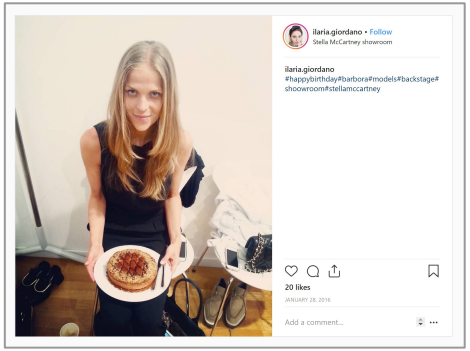
Image Visual Characteristics

Selfie	BodySnap	Marketing	ProductOnly	NonFashion	Face	Logo	BrandLogo	Smile	Outdoor	NumberOfPeople	NumberOfFashionProduct	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
0.026913	0.882489	0.019678	0.003889	0.006827	0.928487	0.209327	0.018385	0.332954	0.021052	1.07905	3.44571	2.36E-05	0.000447	3.79E-06	3.84E-07	0.992289	0.007209	6.17E-06	2.18E-05

Building Strategy: Test Case

- Assume DESIGNER company marketing team performed network analysis, and found potential influencers
- Which to choose?

Option: 1



Predicted: Designer (1)

Actual: Designer (1), Stella McCartney
UserId:1172309667536671450_1300829579

Option: 2

Predicted: High Street (0)

Actual: HighStreet (0), Zara
UserId:1171568394103443082_18001810

Option: 3

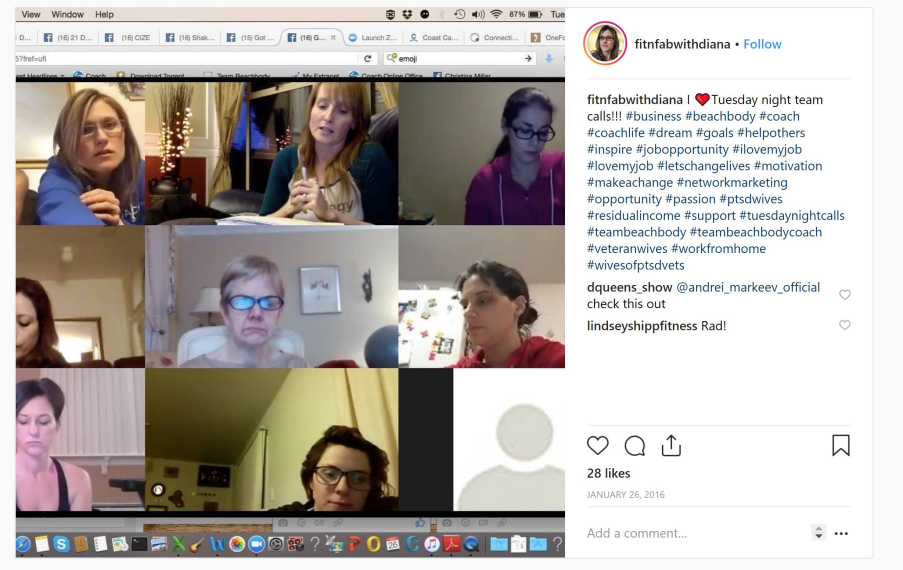
Predicted: Small Couture (2)

Actual: Small Couture (2), Birioni
UserId:1166192422184144257_194690626

Addressing: Incorrect Predictions

UserId:
11714233149723226
76_1500160585

Designer:
Coach



Predicted:
Mega Couture

Actual:
Designer

Image Visual Characteristics

Selfie	BodySnap	Marketing	ProductOnly	NonFashion	Face	Logo	BrandLogo	Smile	Outdoor	NumberOfPeople	NumberOfFashionProduct	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
0.026913	0.882489	0.019678	0.003889	0.006827	0.928487	0.209327	0.018385	0.332954	0.021052	1.07905	3.44571	2.36E-05	0.000447	3.79E-06	3.84E-07	0.992289	0.007209	6.17E-06	2.18E-05

Conclusion and Future Direction

- Benefits:
  - Our model allows us to see which visual features are most associated with each brand category
  - Brands can ensure that their sponsored posts of can utilize similar visual features When choosing among instagram influencers with similar network characteristics

- Limitations & Future Work:
  - Categorization on a brand level (not brand category)
  - Incorporation of additional features that contribute
  - Addressing images with weak brand detection and irrelevant images

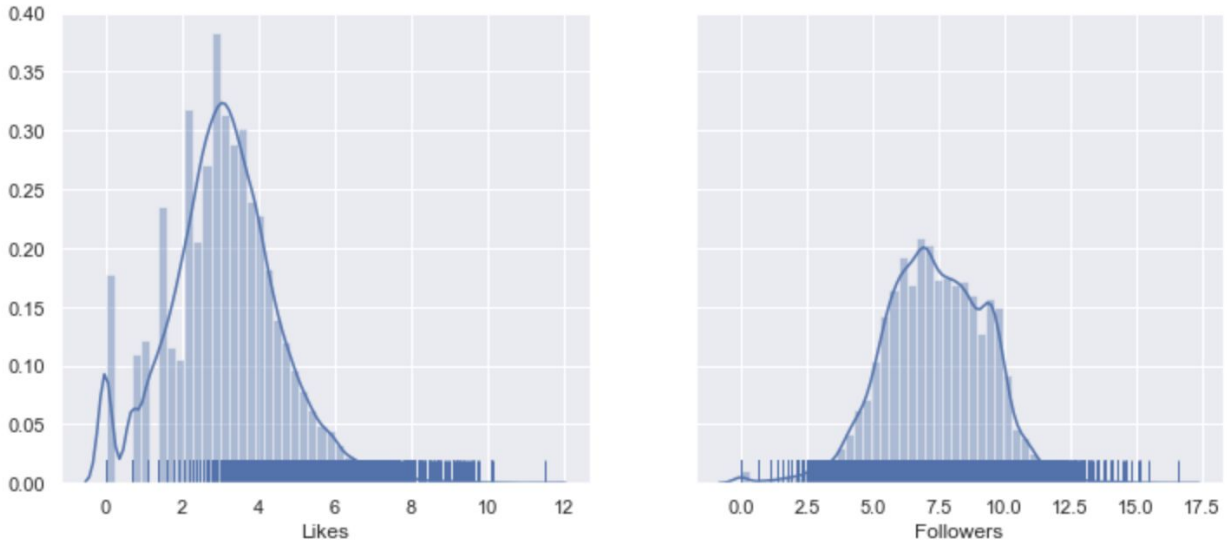
Thank You

Bibliography

- “Fashion Conversation Data on Instagram”, Yu-i Ha, Sejeong Kwon, Meeyoung Cha, Jungseock Joo,ICWSM, 2017
- Hyperparameter Tuning, https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74
- Grid Search Vs Randomized Search,, https://www.analyticsindiamag.com/why-is-random-search-better-than-grid-search-for-machine-learning/

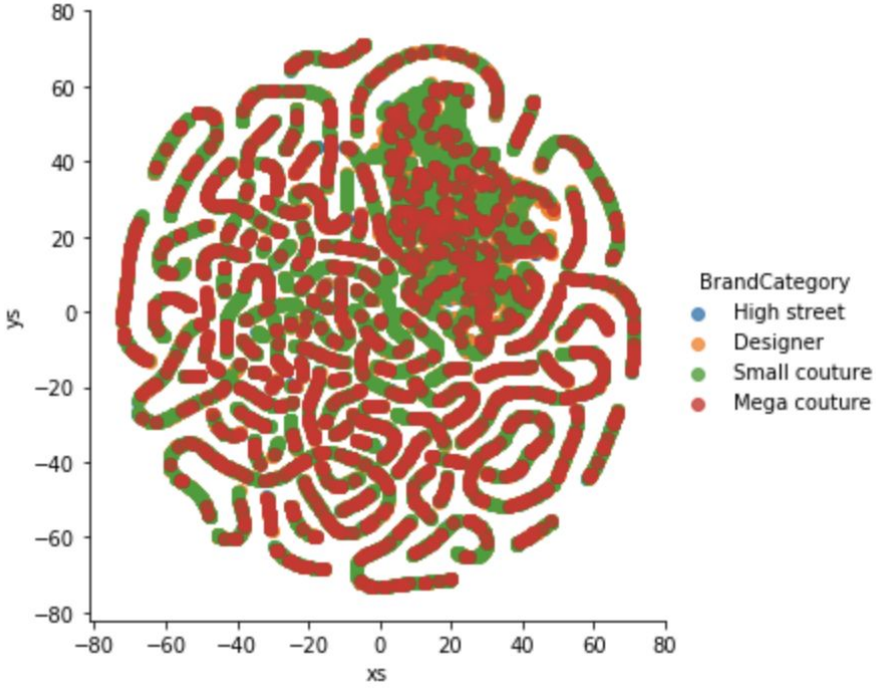
Appendix I: Exploratory Statistics

Log frequencies of number of likes and comments for the posts we have information on



Due to the heavily skewed nature of the data, the log transformation of both likes and followers generates a histogram that is more discernible

Appendix II: Unsupervised Methods: t-SNE



Plotting the brand categories after PCA and t-SNE transformation

Appendix III: Full Prediction List

	Model	Score
4	Random Forest	0.704570
0	Support Vector Machines	0.686292
2	KNN	0.618176
3	Decision Tree	0.598054
1	Linear SVC	0.457372

Appendix IV: Ensemble: Hard Voting with CV

- So far, Random Forest and Support Vector Machine Classifiers performed better than other methodologies
- Fit a hard voting classifier combining RF and SVM

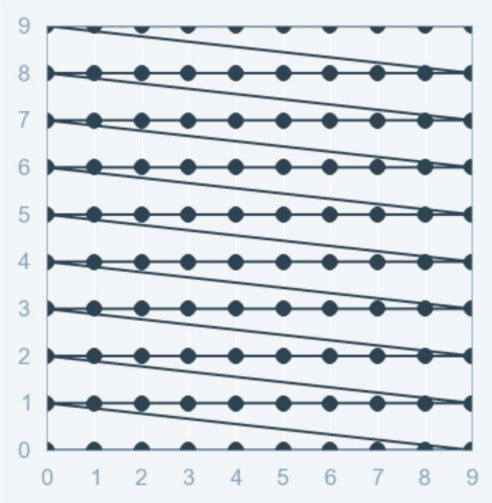
Model Accuracy Score:

0.671

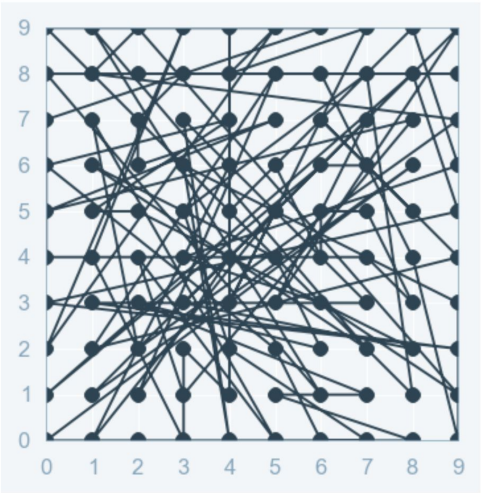
Appendix V: Define Baseline Parameters in Parameter Grid for RF

```
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
```

Appendix VI: Grid Search Vs. Randomized Search Patterns



Grid Search



Randomized Search

Appendix VII: Supervised Methods: Complementary Trivial Solution

Using brand logo variable we can validate, and if necessary, override our model

- This variable is excluded from our model that contains visual features to predict brand category
- If we are over 90% confident a brand logo is present, we can potentially bucket the post into the appropriate brand category
- Using this in conjunction with our model will lead to more realistic outcome