1147 Couse Project

RentHop Inquiry Prediction

Kang Rong

Data Introduction

From Kaggle: 49352 rental listings in New York City in 2016

Numerical Features (5): # of bed/bathrooms, lat/longitude, price

Categorical Features (7): create time, manager ID, street address

Text Features (2) : descriptions (natural language), features (phrases)

Image Feature (1): links of photos

Target Variable (1): interest level ('interested' , 'not_interested')

Missing Values:

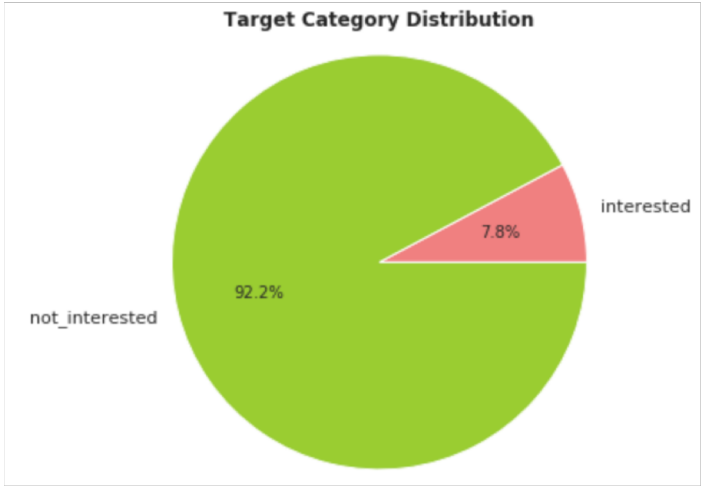
- descriptions / features / photos
- latitude / longitude / street address

Project Goal

Supervised Learning: predict the level of inquires a list will receive

Similar to CTR (click-through rating) prediction

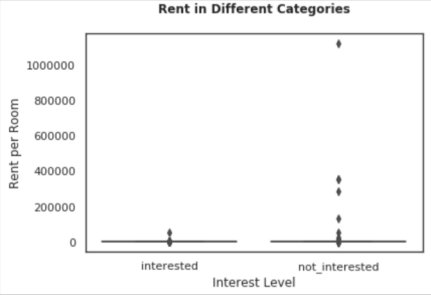
Help e-commercial companies increase sales through ads

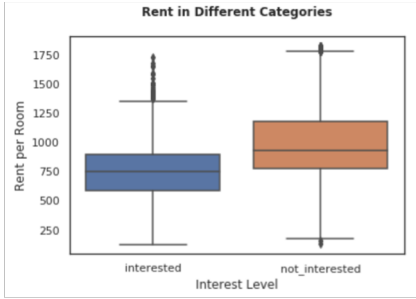


Highly Imbalanced !!!

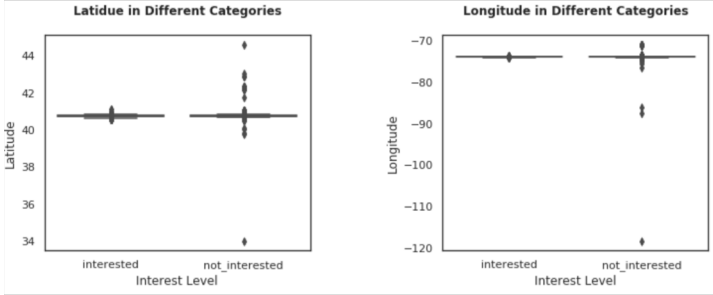
Remove Outliers

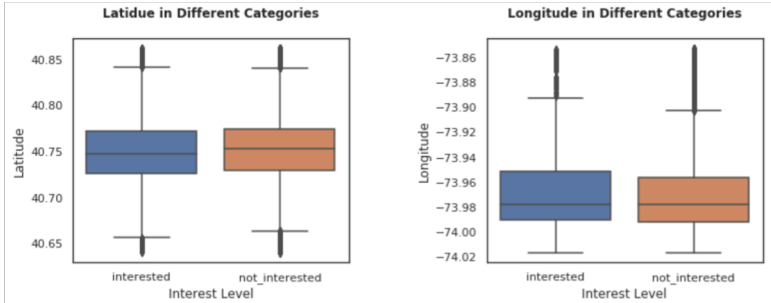
Outliers in prices





Outliers in latitude / longitude





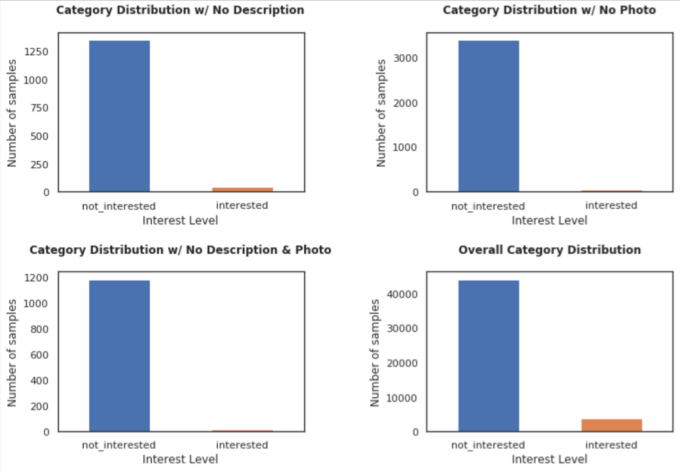
Missing Data

Missing data in street address / latitude / longitude

- Only 12 missing street addresses (drop them!)
- Convert address to coordinates (geolocator API in Python)

Missing data in descriptions / features / photos

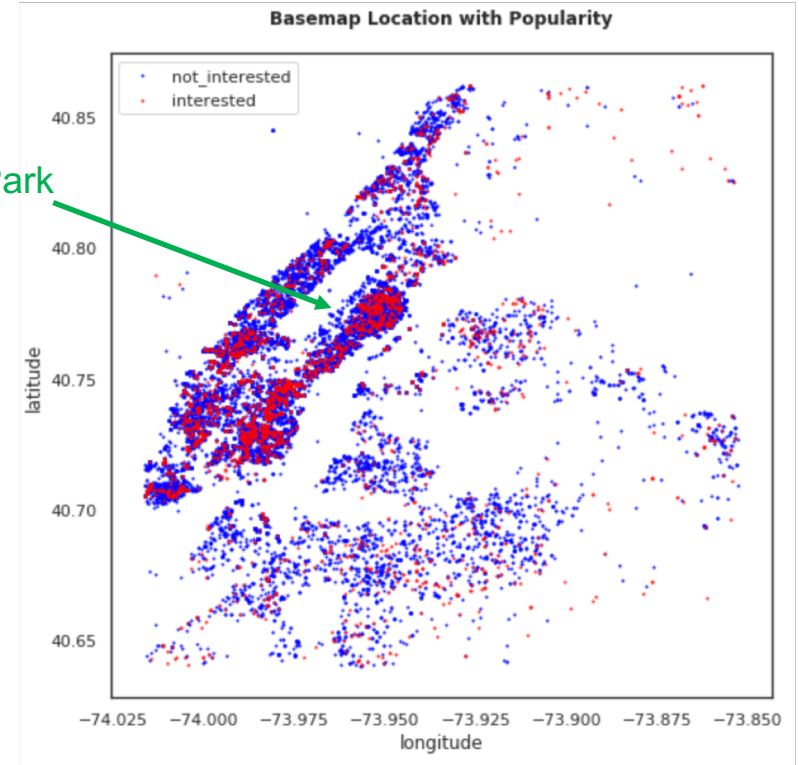
Are they really missing values or do they mean anything?



Listings with no descriptions and/or photos are very likely to be ignored by users !

Analysis of Locations

Display the geographical distribution of rental places



Clustering

Latitude / longitude is not a good feature (numerical ?)

Street address is not good either (too many / no zip code)

Cluster places based on coordinates (K-Means)

Upper Manhattan

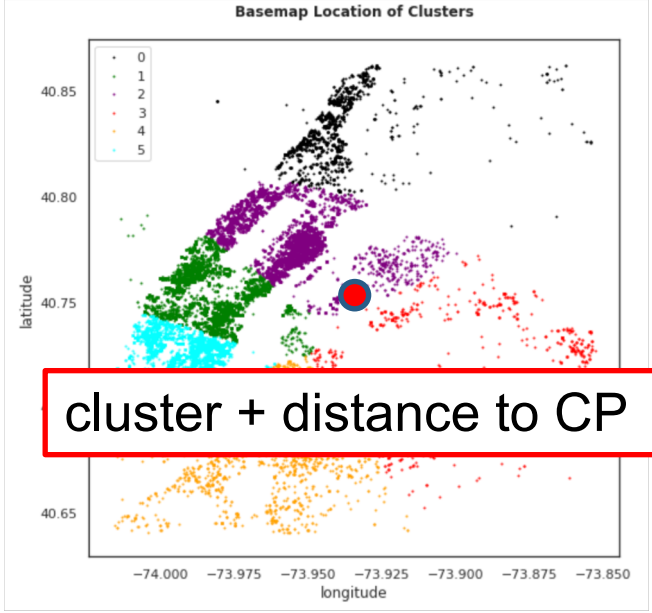
Central Park

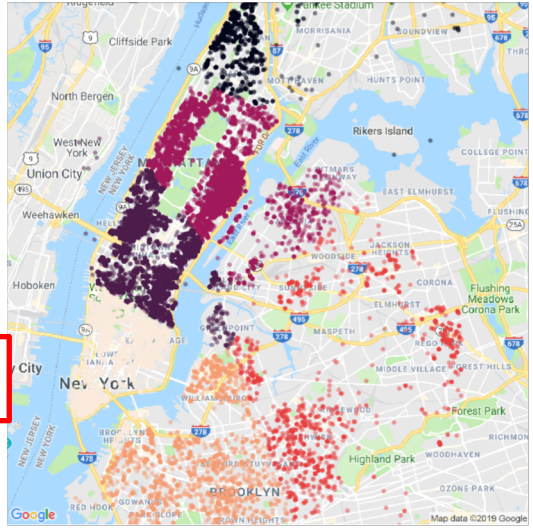
Midtown Manhattan

Lower Manhattan

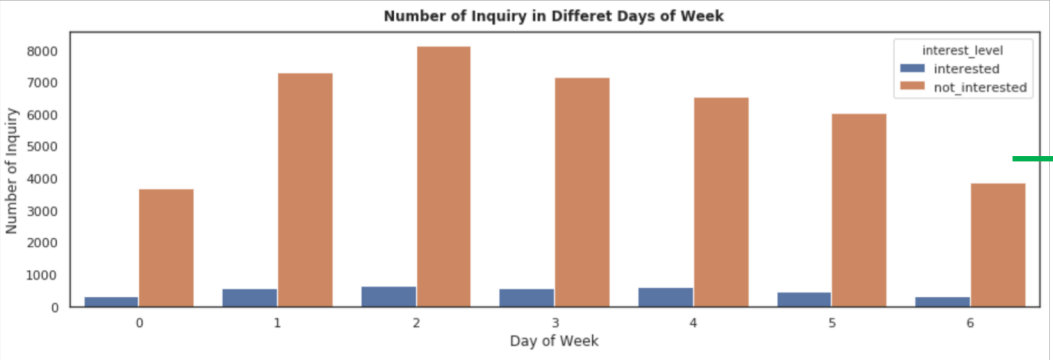
Brooklyn

Queens





Analysis of Created Time

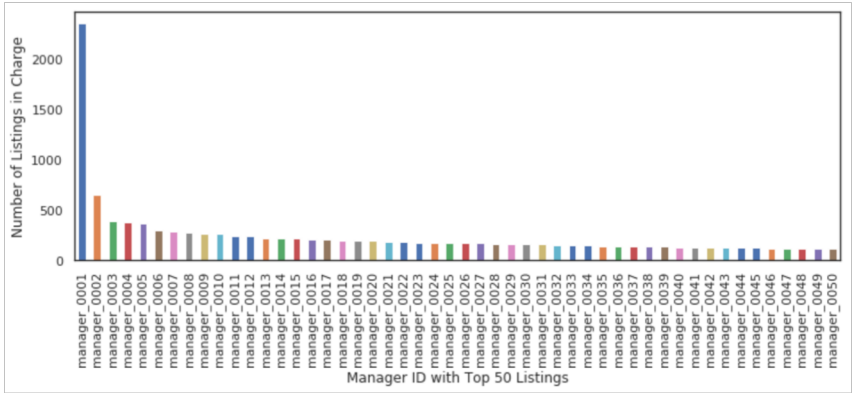


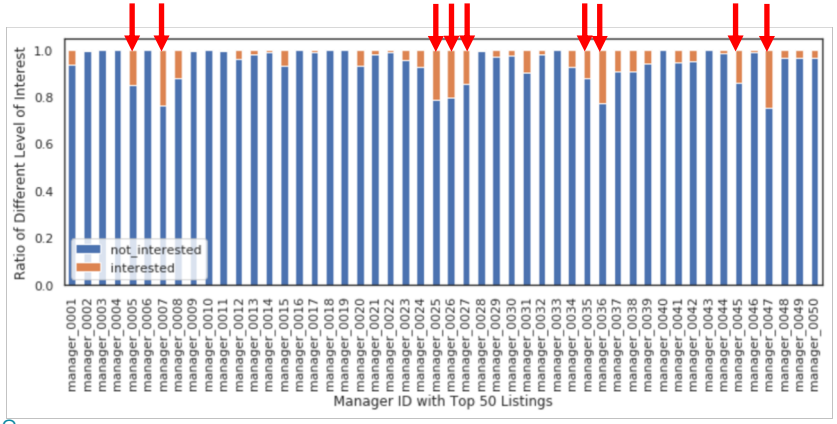
Consistent

Time factor: most of the listing are posted between midnight and early morning

Large proportion of high inquiries

Analysis of Manager Skills





Top 50 managers with the most listings in charge

Manager skills: some managers have higher ratio of highly inquired listings

Text Preprocessing and Analysis

Preprocessing of descriptions:

- Lower case, remove URLs, HTML tags, Email addresses, phone numbers
- Remove punctuations, stop words, meaningless words (contact, call/text, appointment, etc.), word stemming and tokenization.


```
"****AVAILABLE NOW****NO BROKERS FEE****This stunning 1 bedroom (flex 2 bedrooms) 1 bathroom features polished hard-w
ood floors a super-sized kitchen with stainless steel appliances, modern bathroom is decorated with floor-to-wall mar
ble and a mirrored medicine cabinet and finishing touches, living room can fit three single couches and a coffee tabl
e, bedroom can fit a queen-size bed and a side table and has an abundant of closet space.-- Floor-plan attached***10
0%-NO FEE***<br /><br />*****LOCATION BENEFITS*****Located in the heart of Manhattan in a thriving, diverse neighborh
ood, with a variety of shops,restaurants, bars,lounge's,live theater, cinema, and even alocal Comedy Club, in walking
distance.******BUILDING AMENITIES******?24 HR DOORMAN?ELEVATOR?LAUNDRY IN BUILDING?GARAGE PARKING?HEALTH CLUB (FREE G
YM MEMBERSHIP)?SUN-DECK?PACKAGE SERVICES?LANDSCAPED OUTDOOR SPACE?DRY CLEANING SERVICE<br /><br />*****PUBLIC TRANSPO
RTATION******---  E 28th st Subway: 6 Train,---  Cross-Town Buses.<br /><br />FOR FURTHER INFO AND TO SCHEDULE A VIEW
INGCONTACT: NASH BENAIM  CALL/TEXT: 304-782-4563  EMAIL: kagglemanager@renthop.com<p><a  website_redacted "
```



```
'avail no broker fee stun flex bedroom featur polish hard wood floor super size kitchen stainless steel applianc mode
rn decor floor wall marbl mirror medicin cabinet finish touch live fit three singl couch coffe tabl fit queen size be
d side tabl abund closet space floor plan attach no fee locat benefit locat heart manhattan thrive divers neighborhoo
d varieti shop restaur bar loung live theater cinema even aloc comedi club walk distanc amen hr doorman elev laundri
garag park health club free gym membership sun deck packag servic landscap outdoor space dri clean servic public tran
sport th st subway train cross town buse info schedul viewingcontact nash benaim'
```

Text Preprocessing and Analysis

Ideas:

- TF-IDF + Sentiment Analysis? May not be a good idea.
- Generate features (keywords) from descriptions and one-hot encoding

Problems:

- Features with similar/same meanings

Solutions:

- Replace similar features with the an uniform expression
- ~1000 features → ~400 features

cats / dogs allowed ; pets allowed ; pets ok ⇒ pet

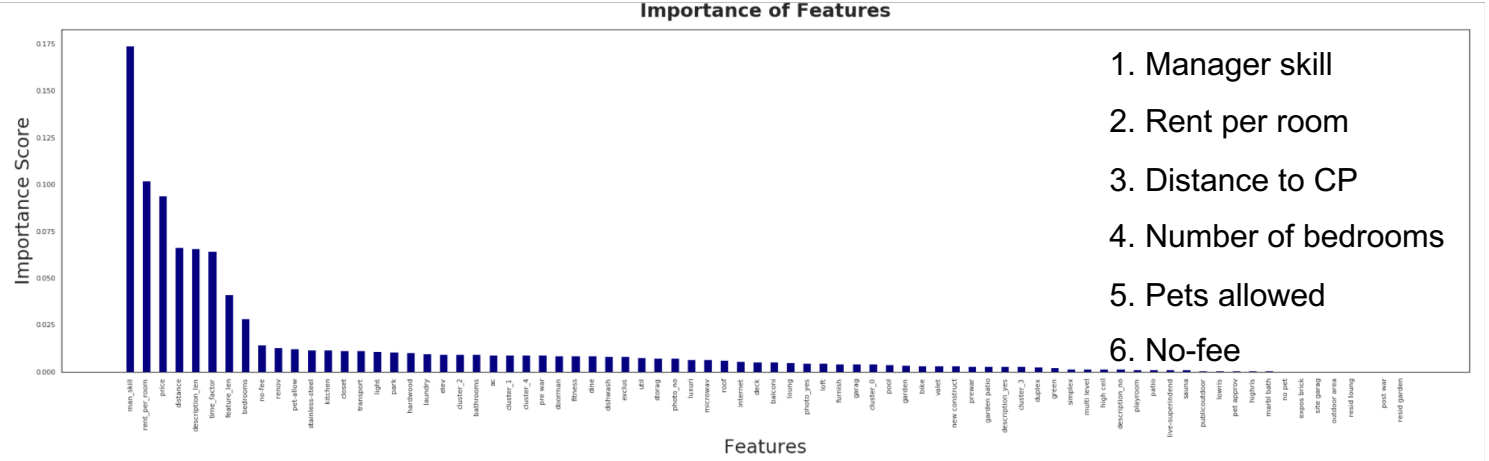
washer / dryer in unit ; laundry in unit ; lndry in building ⇒ laundry

stainless steel; stainless; ss ⇒ stainless

Feature Engineering

Feature Generation and Selection

- Add: rent per room, description length, feature length, w/wo photo (1,0)
- Add: manager skills score 0 x (not_interested_ratio) + 3 x (interested_ratio)
- Add: time factor score [5-7 am] x 3, [1-5 am] x 1, otherwise x 0
- one-hot encoding (100 features in total) and data standardization



Challenge: Imbalanced Classes

Subsampling: roughly 1:1 of majority and minority classes (info loss)

Oversampling: Duplicate samples in minority class (overfitting)

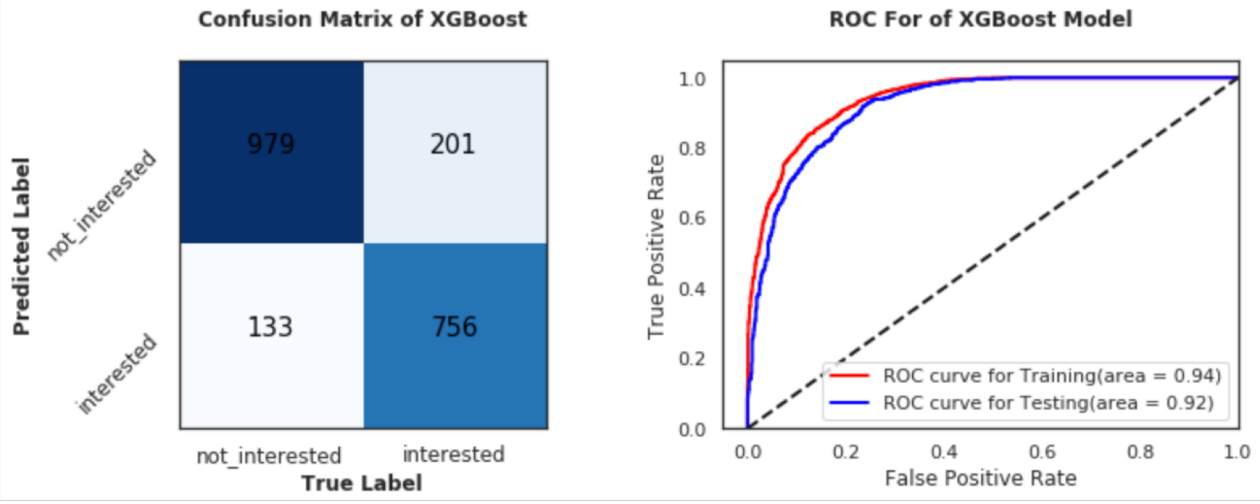
Random Forest: class_weight = 'balanced'

XGBoost: recursively fit the error model

Base Model with Subsampling

Subsample (1:1) and apply base model of XGBoost

- Subsampling achieves the goal of solving data imbalance issue
- ROC shows descent result

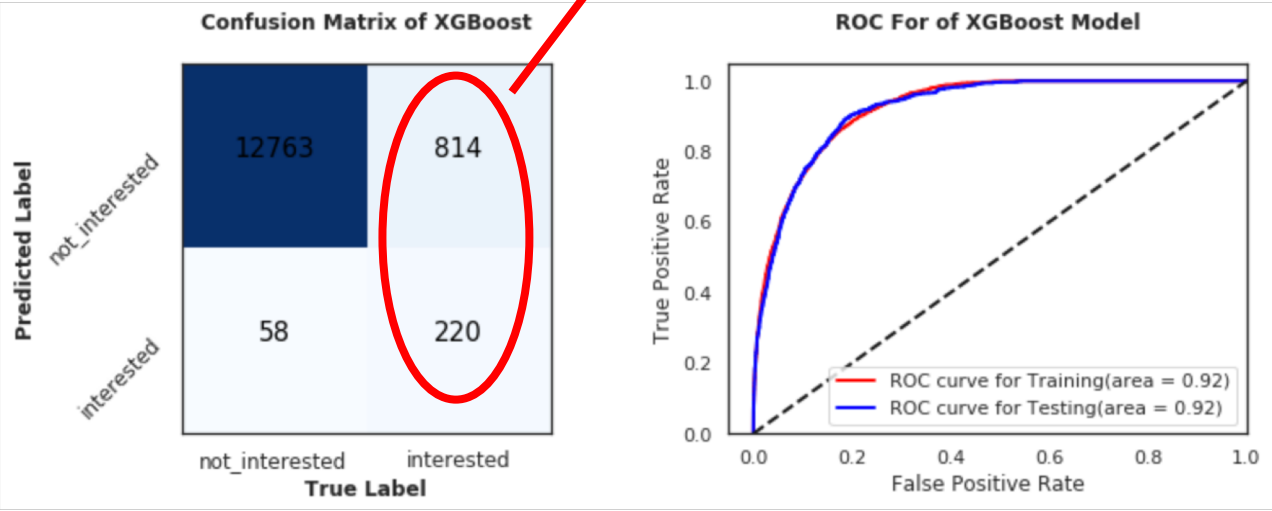


Base Model with Original Data

Still attempts to use the original data

- Subsampling has the risk of losing information

Imbalance issue not solved

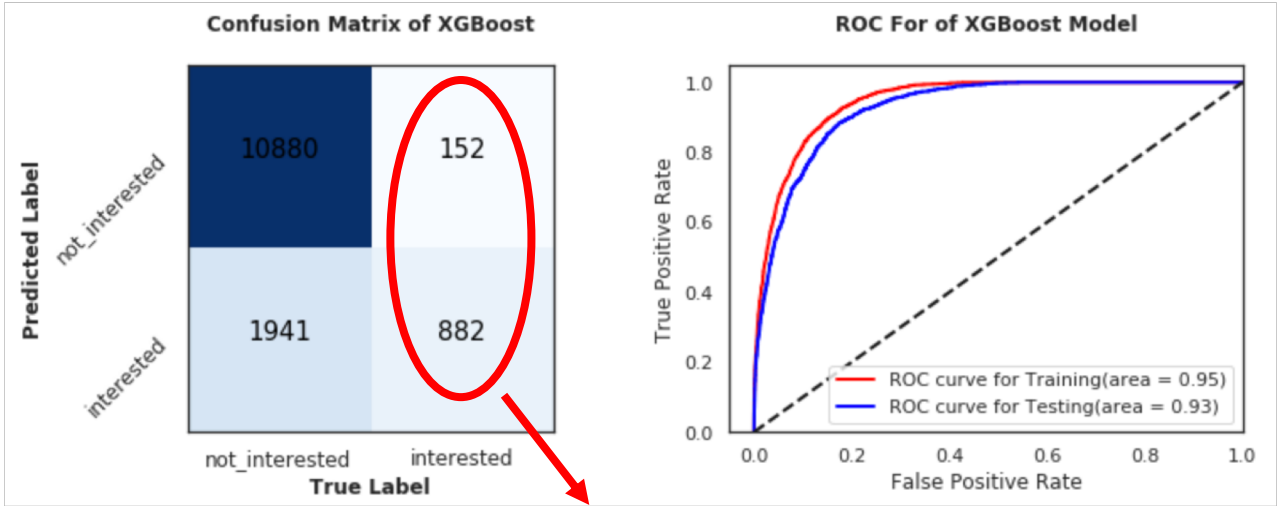


Hyperparameter Tuning

Two important hyperparameters to deal with imbalanced data:

- **scale_pos_weight**
- **max_delta_step**

scale_pos_weight = 10
Max_delta_step = 5



Imbalance issue solved

Future Improvements

Feature Engineering: find some more potential features

Image Data: find a better way of utilizing the image data

Neural Networks: Try DNN to improve model performance

THANK YOU!