Image Classification for Property Tech

Capstone Project for General Assembly Data Science Immersive Program (DSI 19, Singapore)

By Stephen Chan

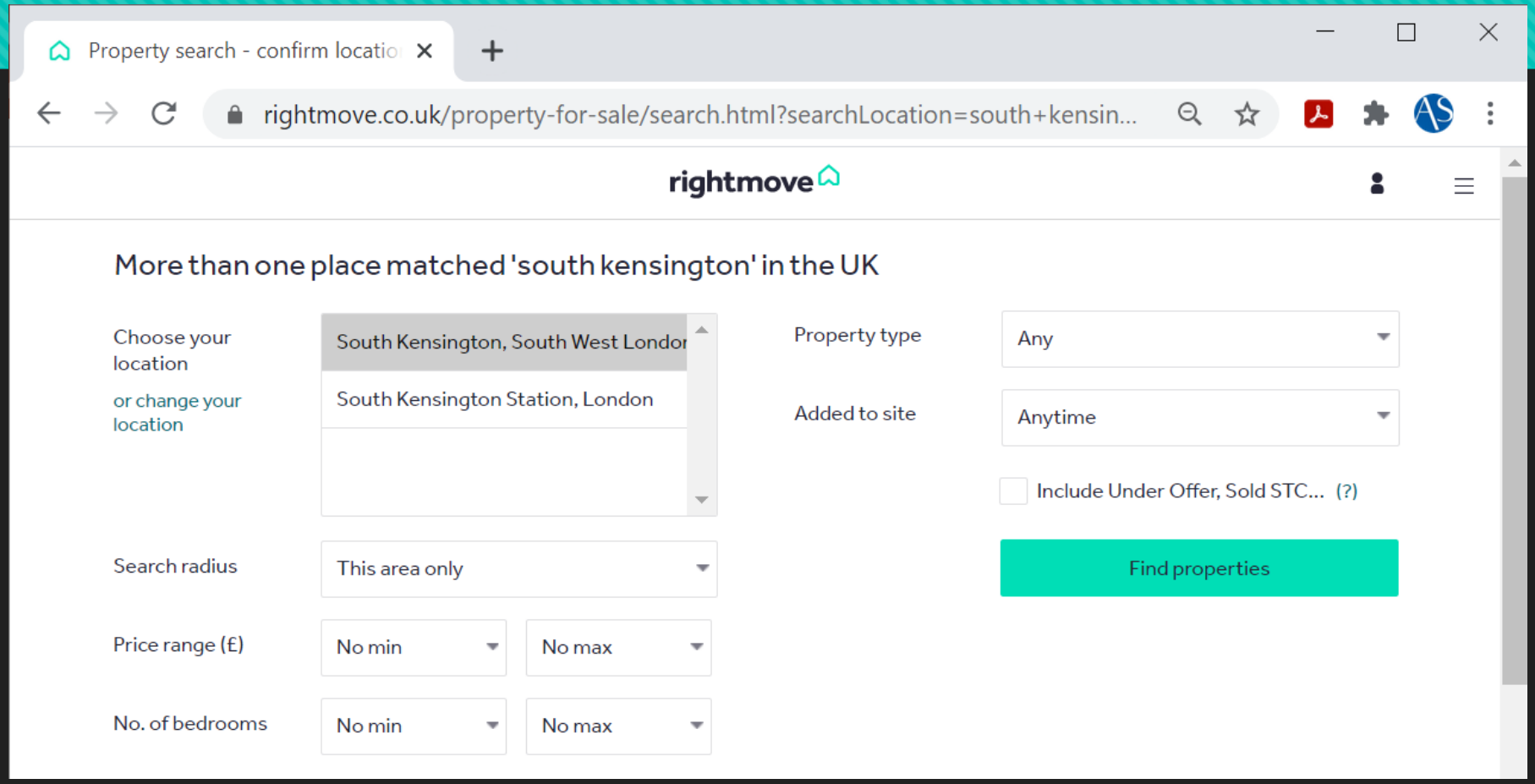
26 February 2021

Problem Statement

Property Tech

- London property information is sporadic and disorganised
- One crucial piece of information is not easy to obtain in the public domain
  - Age of Building, or simply categorised as Period/Old vs Modern
- Most buyers spend a great deal of time in property search, and receive poor recommendations
- This project aims at filling the information gap
  - by classifying buildings as Period vs Modern, based on 2000+ images from leading portals (Rightmove, Zoopla, Purple Bricks)
- Results: metrics ~90% (AUC, accuracy)

Current Applications for Property Search - Needs Improvement



Example of UK property images

Data Source

- 3000+ images from the public domain
- 2000+ labels done manually

Features to extract

1. Period vs Modern (primary goal)
2. Exterior vs Interior (secondary, which turns out to be crucial)

Period - Exterior

Period - Interior

Modern - Exterior

Modern - Interior

Challenges: Various Exterior Styles

- Period Buildings:
  - Georgian (1714-1837)
  - Victorian (1837-1901)
  - Edwardian (1901-1910)
- Council houses
- Old houses
- Terraced or detached houses





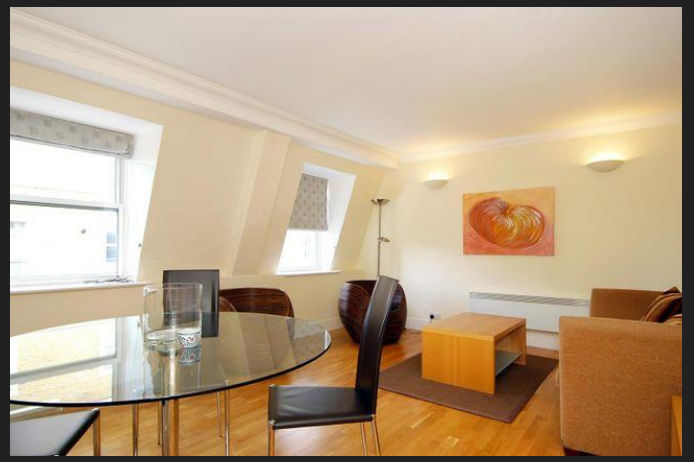




Challenges: Refurnished Interior of Old and Period Buildings

- Easy
  1. Particular type of window frame
  2. Tilted walls
  3. Wood burning stoves
- Difficult
  1. No windows
  2. Ground floor flats with large windows in the living room

Easy



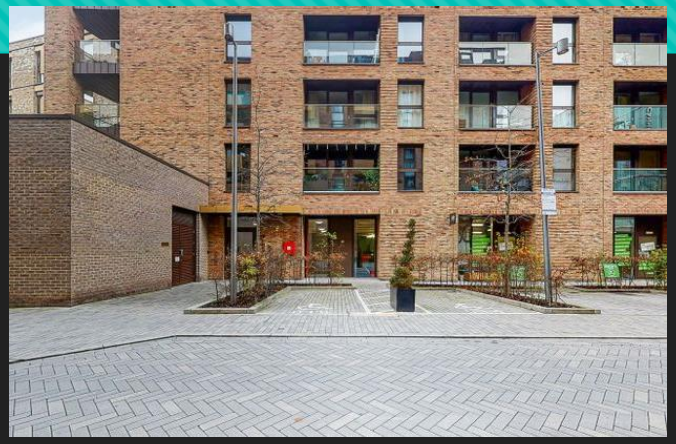
Difficult

Challenges: Modern vs Retro Buildings

- Easy
  1. High-rise condominiums
- Difficult
  1. Low-rise buildings with brick walls

Easy

Difficult



Other Challenges

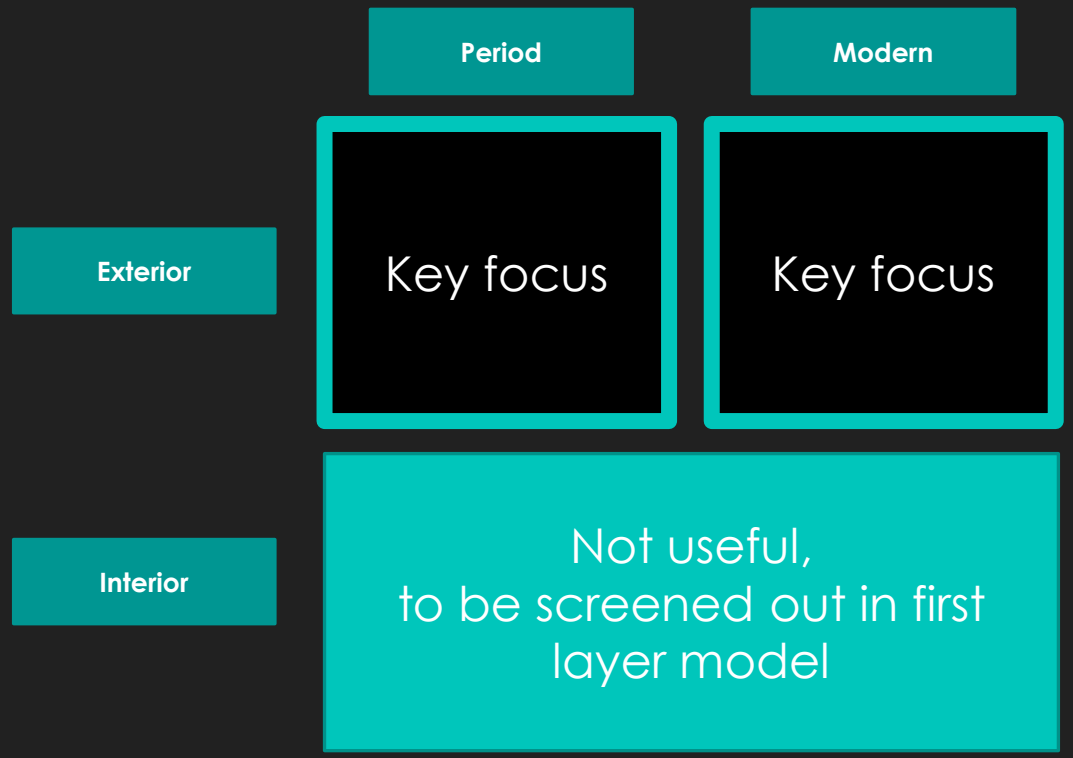
Property Tech

- Other challenges
  1. Blocked views
  2. Watermarks
  3. Repeated images

Watermarks

Blocked Views

Divide and Conquer



- The ultimate goal is to classify period vs modern buildings
- Limitation on accuracy based on interior images
  - Sometimes classification may not be accurate even by human eyes
- Strategy: Divide and Conquer
  - Separate "good" images from the pool ie identify "exterior" images as they tend to give more information
  - Feed through two classification models
    - 1st layer model for classifying exterior vs interior
    - 2nd layer model for classifying period vs modern, based on exterior images only
- Model comparison
  - Customised CNN vs Transfer Learning (VGG16)
- Result: metrics improved to over 90% (from 70%)

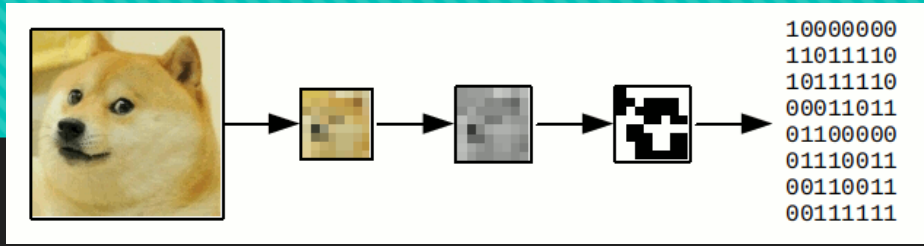
Data Exploration and Preprocessing

Web Scrapping, manual labelling

Property Tech

- Web scrapping of the major property portals in UK
  - Rightmove
  - Zoopla
  - PurpleBricks
- 2000+ images obtained
  - Information gathered:
    - one image from each property listing
    - location information eg coordinates
    - Description
- Manual labelling
  - Basic key word search eg Georgian , Victorian etc to do first level of labelling
  - Glancing through the images as second level of labelling

Image Hashing, Resizing, Standard Scaling



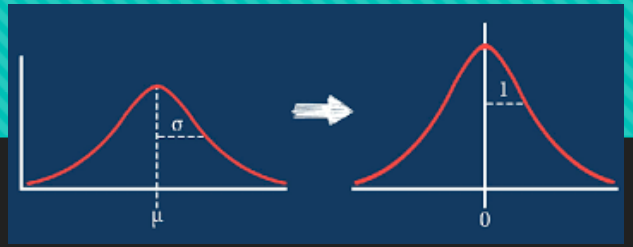


Image Hashing

- ImageHash* was deployed to eliminate duplicates
- Hashing is a function that applies to an arbitrary data and produces the data of a fixed size
- Average hashing algorithm:
  - Scale the image
  - Convert to greyscale
  - Calculate the mean and binarize the greyscale based on the mean
  - Convert the binary image into the integer
- By trial and error, hash size of 8 is appropriate. Below 8 will cause too many different images under the same hash
- Duplicated images in the entire image dataset are identified. Duplication level around 6%

Image Resizing

- All images are resized to 300 x 300 px for modelling fitting
- Higher resolution will cause kernel instability

Standard Scaling

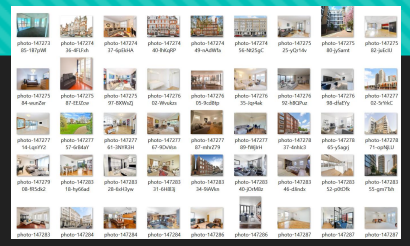
- Standard scaling is applied to each channel of a pixel across the training set images (not the test set)

Notes:

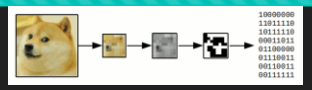
https://github.com/JohannesBuchner/imagehash

Overview of Data Preprocessing and Modelling

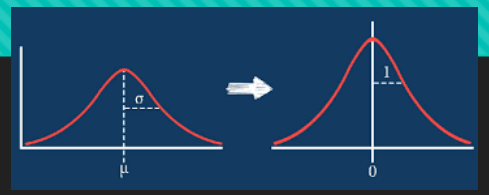
Images (Original)



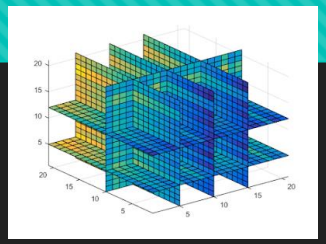
Duplicates removed after hashing

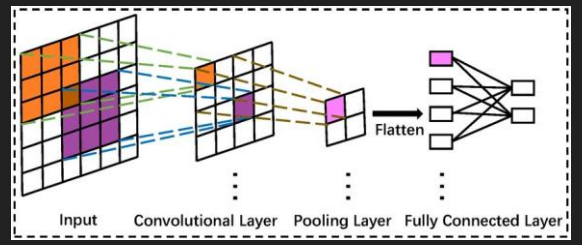


Standard Scaling

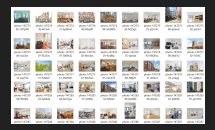


Data in arrays

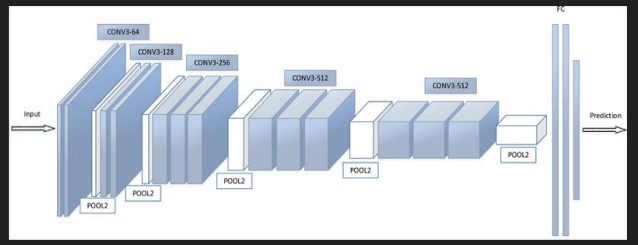




1st layer model - Customised CNN

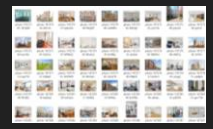


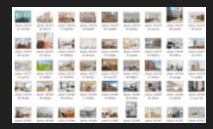
Exterior images



2nd layer model – VGG16

Exterior - Period





Exterior - Modern

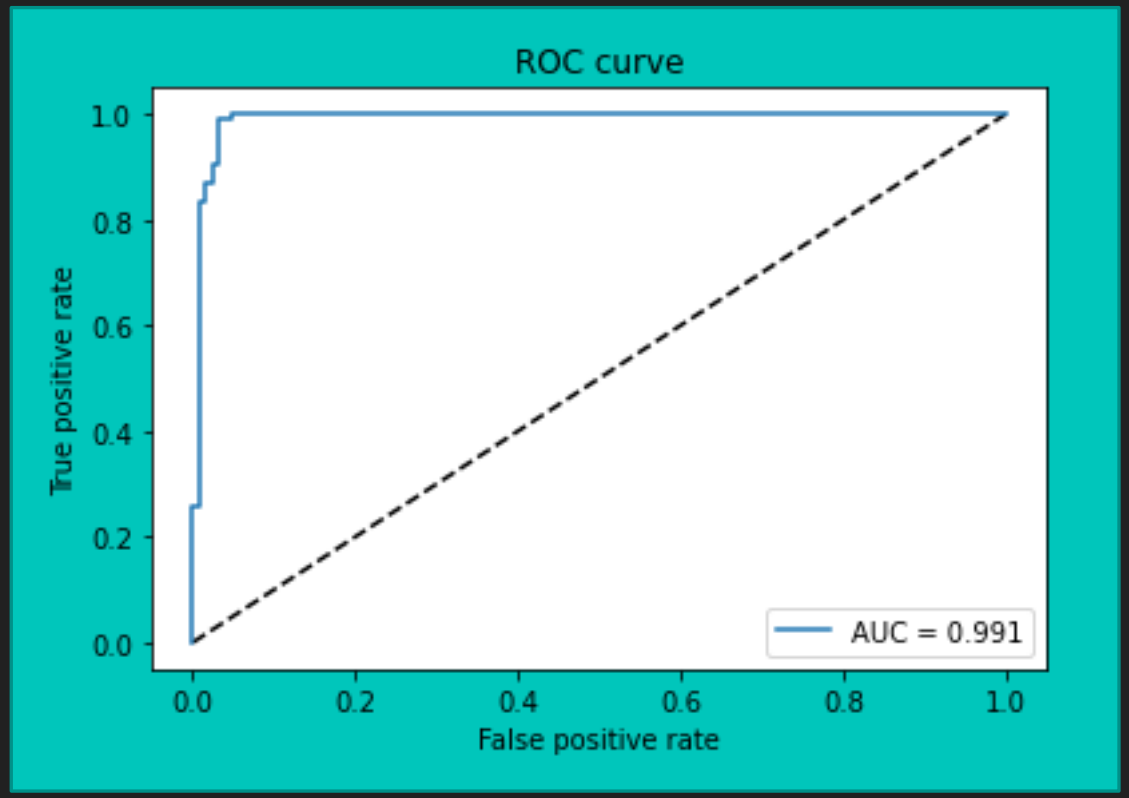
Modelling – CNN and Transfer Learning

- Customised Convolutional Neural Networks (CNN)
  - Customised structure
    - Two convolutional layers with 6 and 16 filters respectively, each followed by a max pooling layer of 9x9
    - One flattening layer
    - One layer of 256 densely connected neurons
    - One dropout layer, dropout rate 0.5
    - Final Sigmoid activation layer
    - Optimizer: Adam
  - Grid Search CV was applied to determine the best parameters
  - Faster computation in each iteration, but more iterations needed

- Transfer Learning based on VGG16
  - VGG16 as the input model
    - No change in the original VGG16 structure and parameters
    - One additional layer of 128 densely connected neurons
    - One dropout layer, dropout rate 0.5
    - Final Sigmoid activation layer
  - Quicker in convergence, less than 15 epochs based on the tolerance
  - Longer computation in each iteration due to many more layers

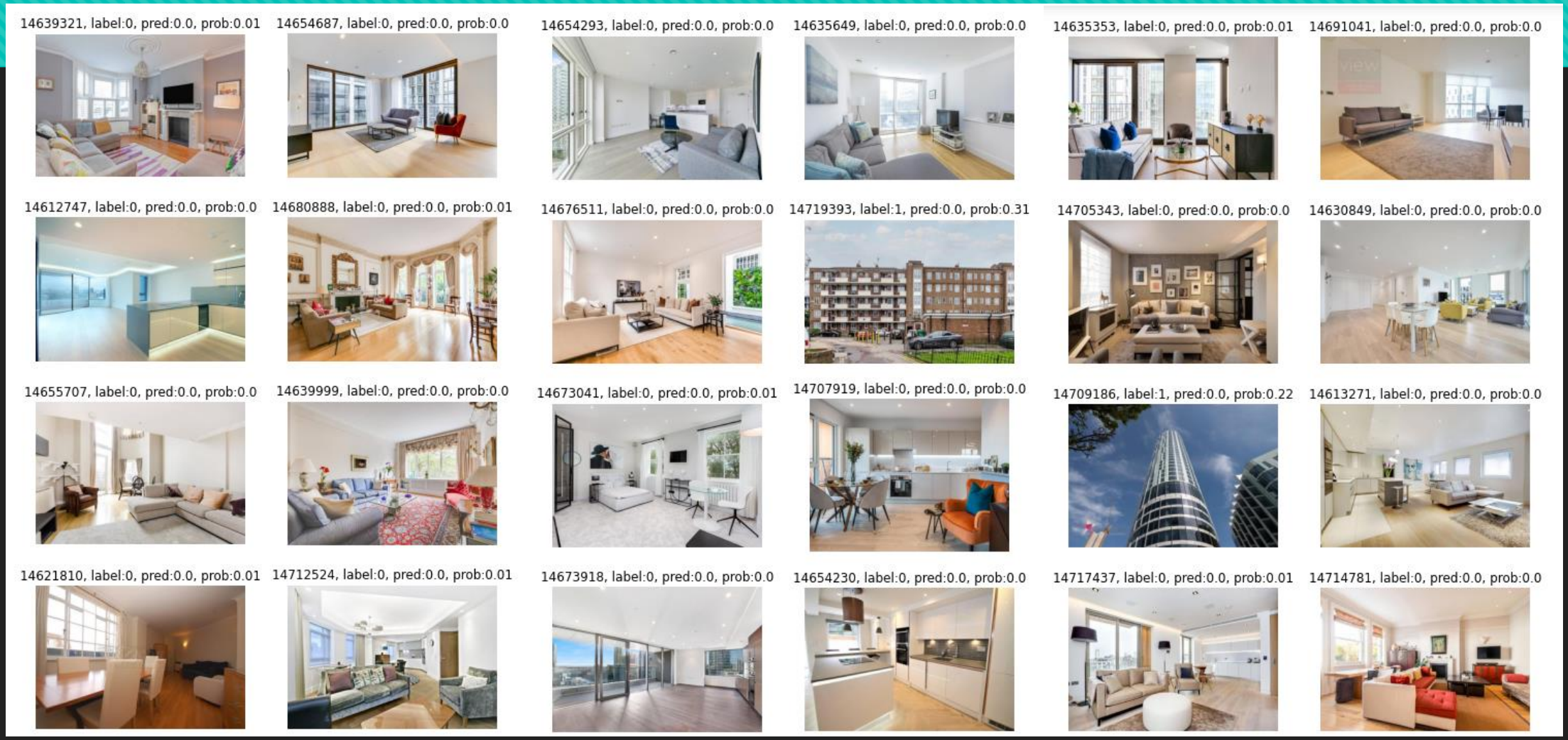
Modelling Results – First Layer - Exterior vs Interior

- Customised CNN
  - Classified images into exterior vs interior very effectively
  - AUC: 99%
  - Accuracy: 95% (238/250 in validation set), vs baseline at 50% (balanced class)
  - Data set of 1000 images, 500 each class (Exterior vs Interior)



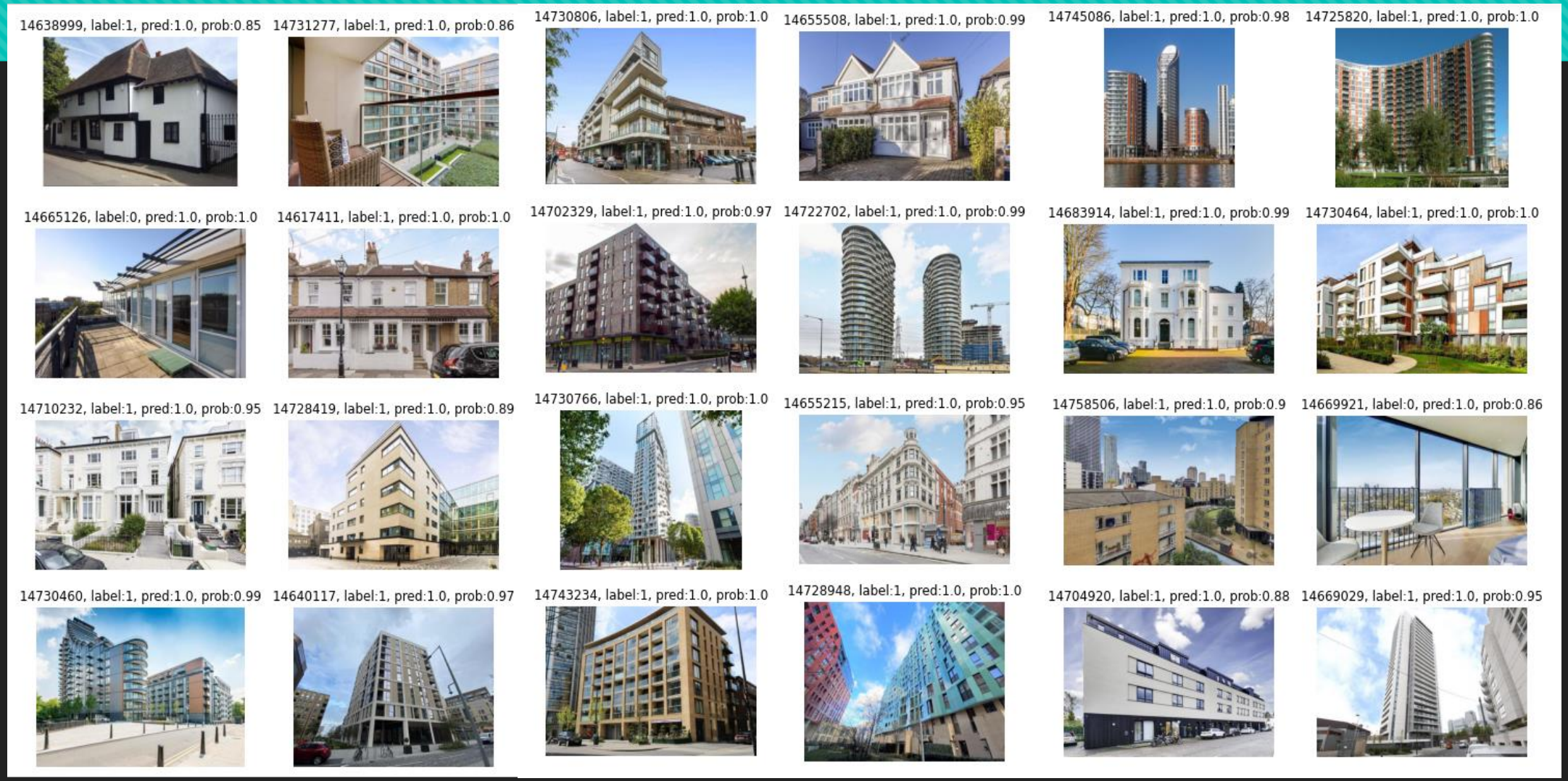
Modelling Results – First Layer - Interior vs Exterior (cont'd)

- Classified as Interior



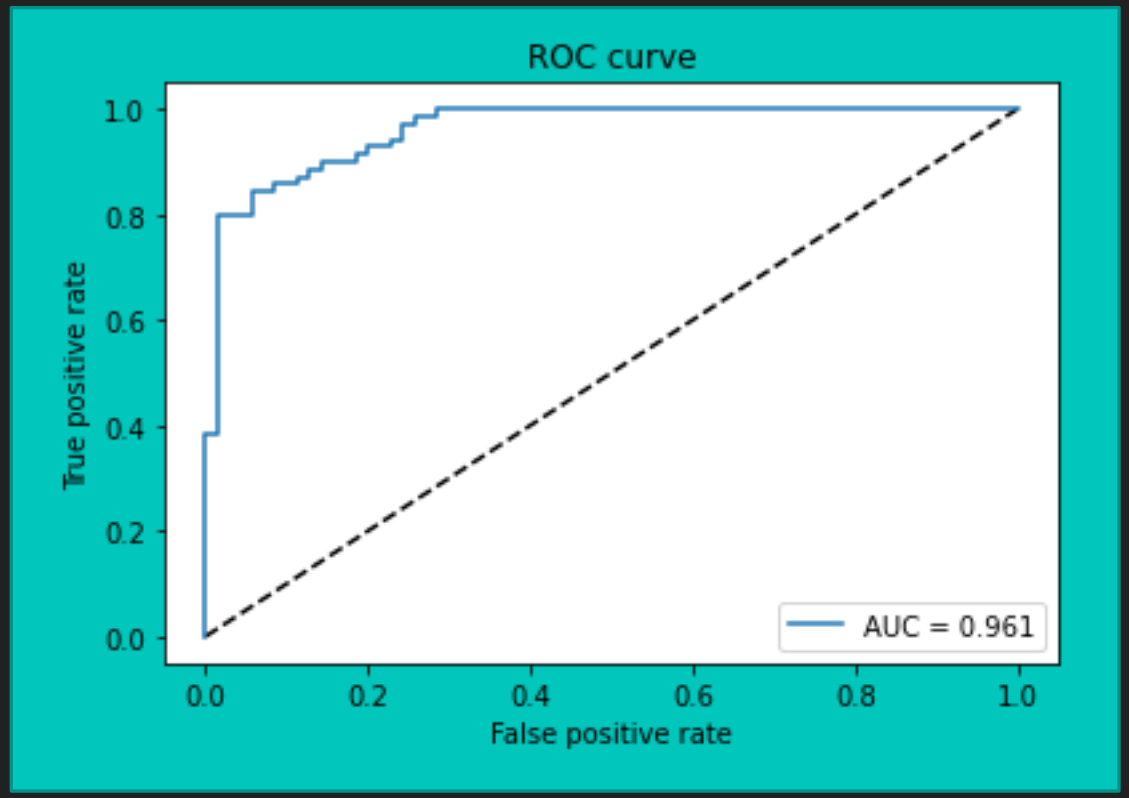
Modelling Results – First Layer - Interior vs Exterior (cont'd)

- Classified as Exterior



Modelling Results – Second Layer - Period vs Modern

- Transfer Learning CNN (VGG16)
  - Classified exterior images into period vs modern very effectively
  - AUC: 96%
  - Accuracy: 88% (123/140 in validation set), vs baseline at 50% (balanced class)
  - Data set of 560 images, 280 each class (Period vs Modern)
  - The transfer learning model is superior to the customised CNN (AUC ~85% and accuracy ~70%)

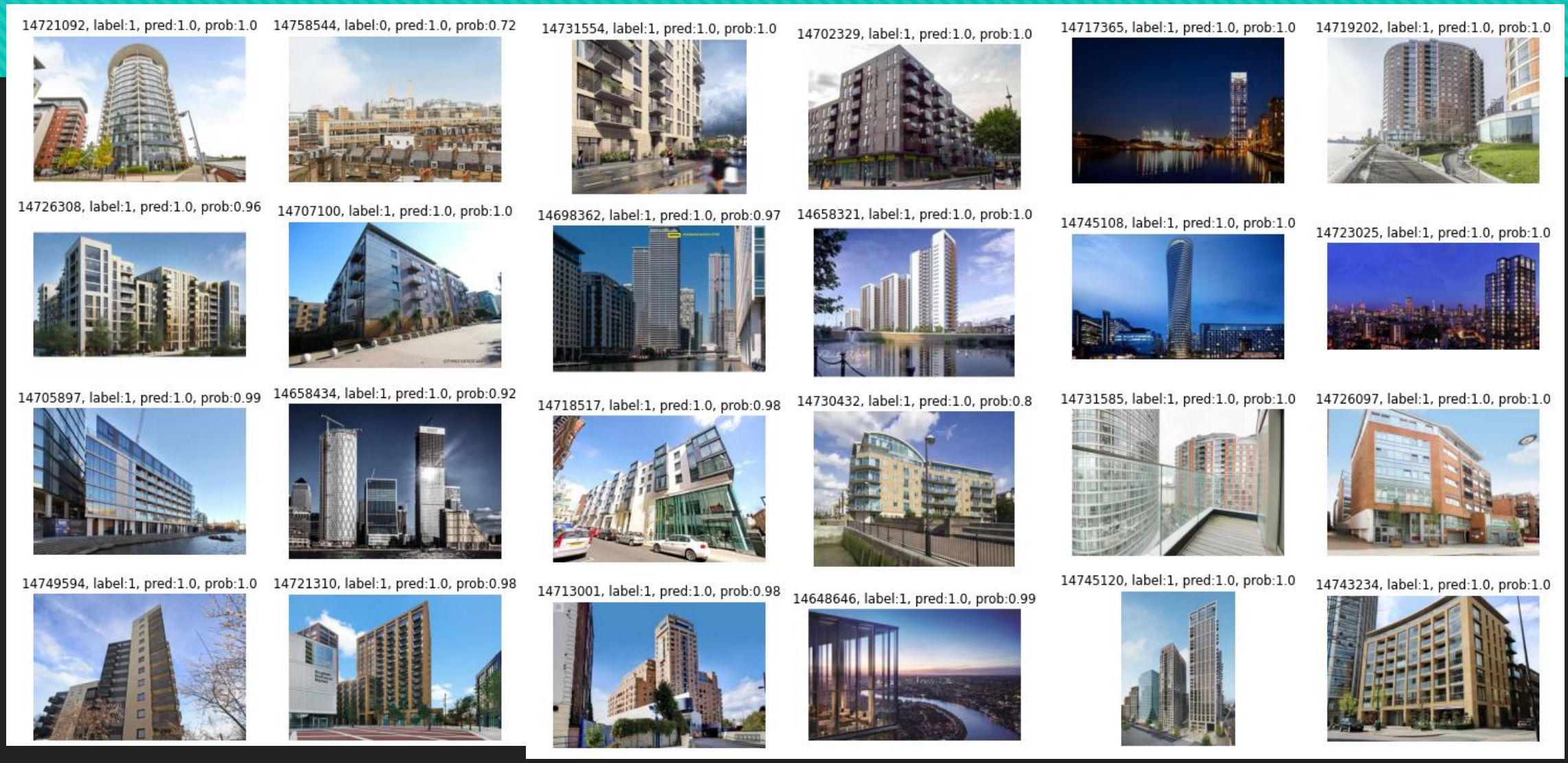


Modelling Results – Second Layer - Period vs Modern (cont'd)

- Classified as Period / Old

Modelling Results – Second Layer - Period vs Modern (cont'd)

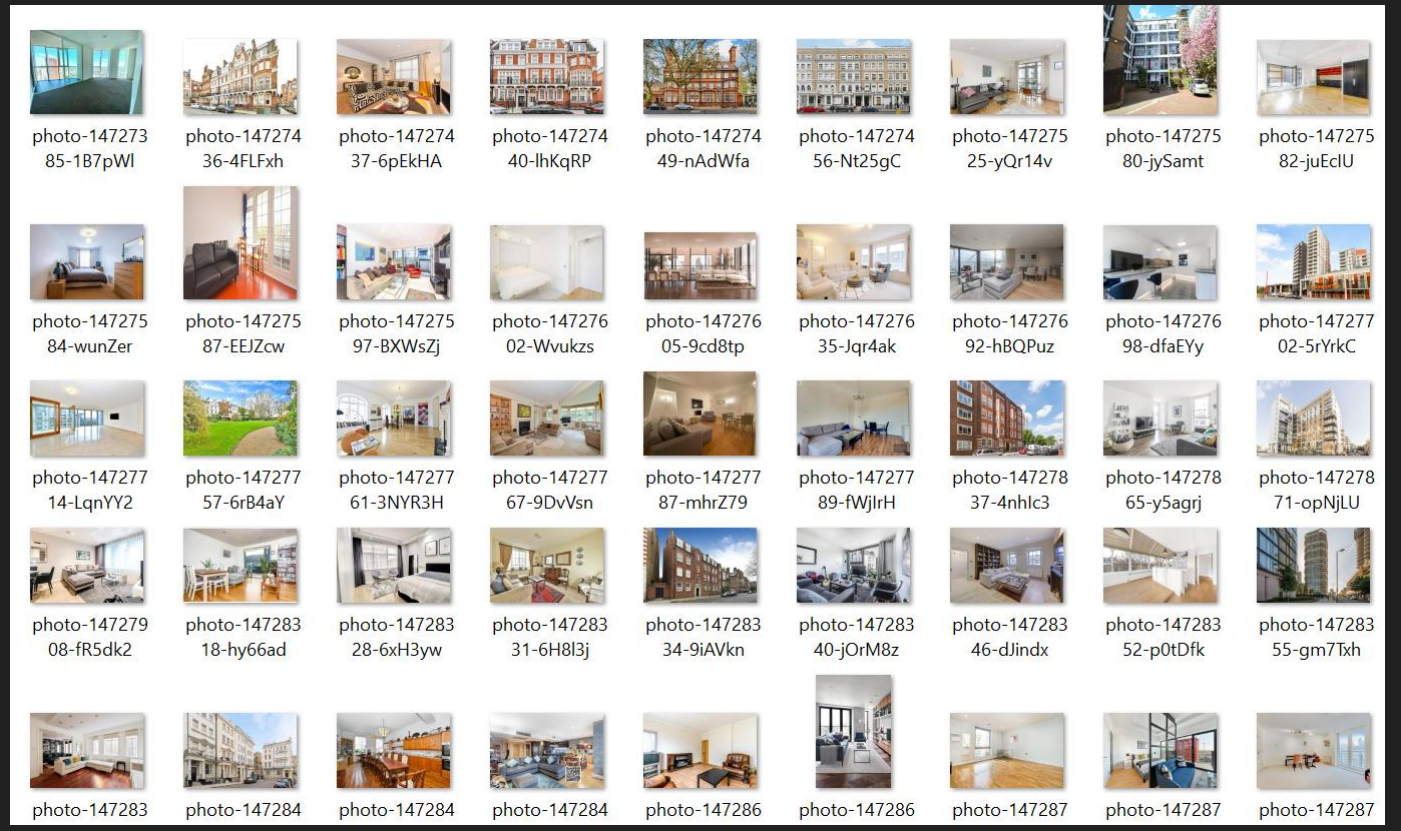
- Classified as Modern



Final Challenge - Unseen Data

Unseen Data without labels

- 1000 images (unseen data) are fed through two layers of model
- No labels
- Model outputs are evaluated by eyeballing the image classes, histogram of predicted probabilities, and plotting the classification on the London map



Unseen Data without labels

- **Results:**

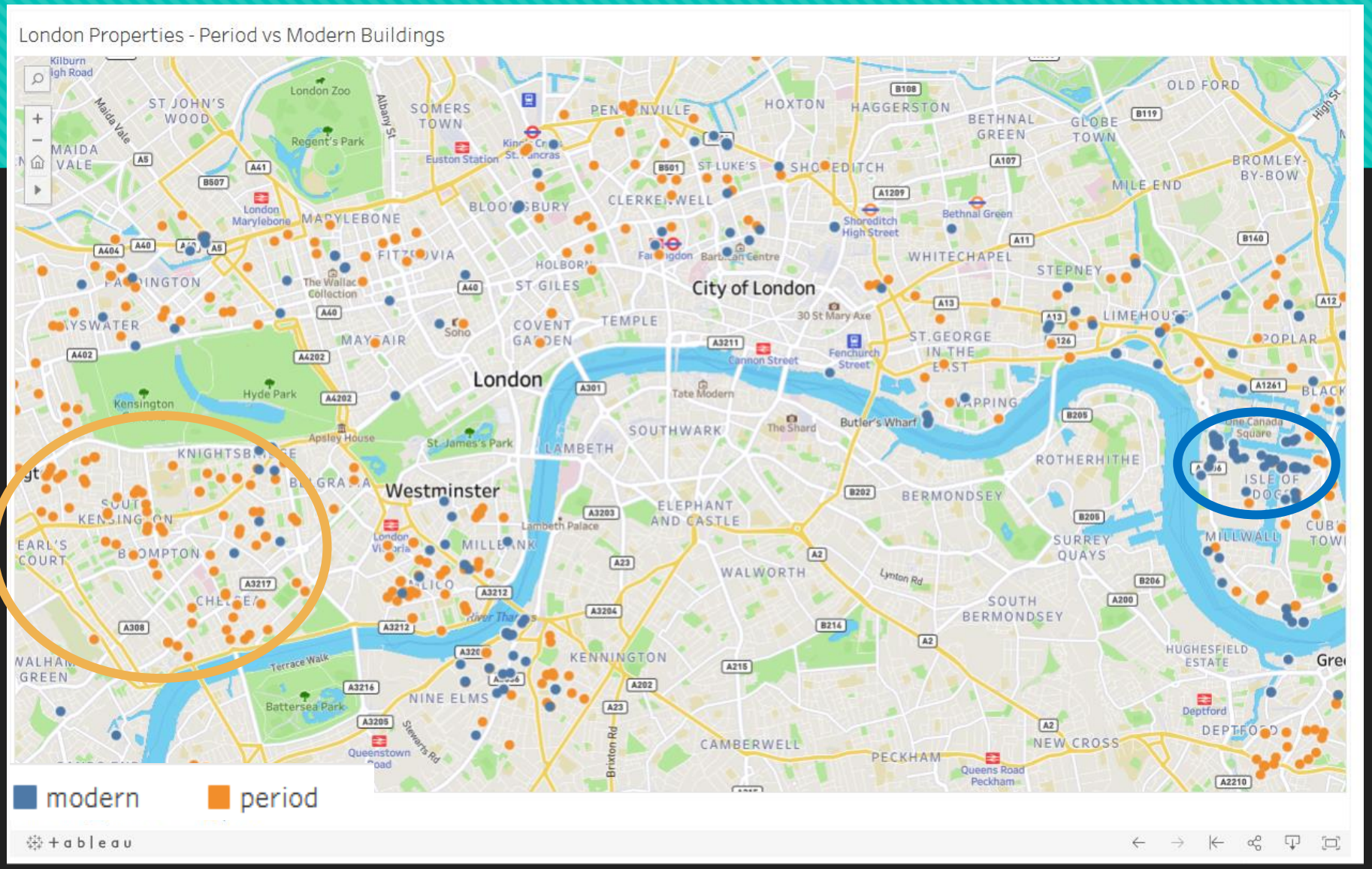
  First layer

  - **Very high accuracy of classifying exterior vs interior, similar to results of training/validation**

  Second layer

  - **Overall high accuracy of classification (period vs modern). Higher accuracy in classifying period buildings**
  - **Map plot is demonstrating right classification in the districts**
    - **South Kensington - primarily period buildings**
    - **Canary Wharf – mostly modern buildings**
    - **Clerkenwell, Bloomsbury, Westminster - mix of period buildings and some modern buildings**

Second Layer Model Results – Map Plot



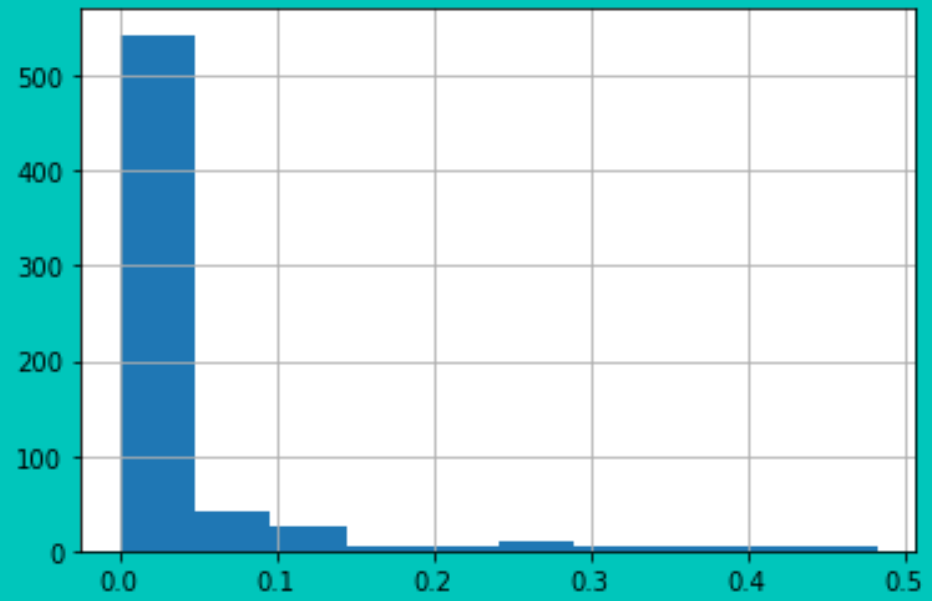
South Kensington is known for period houses (eg Victorian)

Canary Wharf is known for modern condos

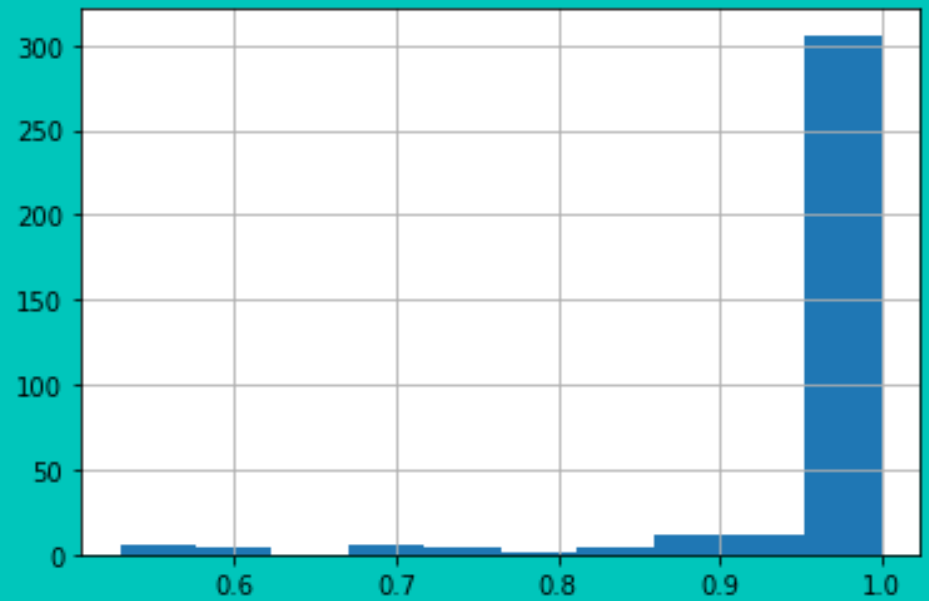
First Layer Model Results – Interior vs Exterior

First Layer Model Results – Histogram of Predicted Probabilities

Unseen images classified as Interior (probability < 0.5)



Unseen images classified as Exterior (probability >= 0.5)

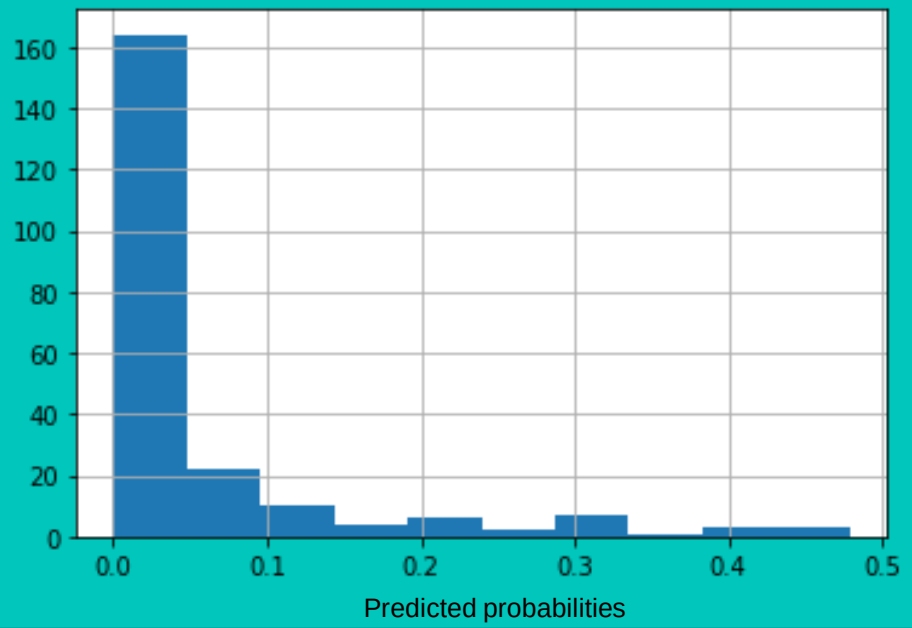


Majority of predicted probabilities are close to 0 and 1, ie high confidence of classification.
Confirmed by human eyes

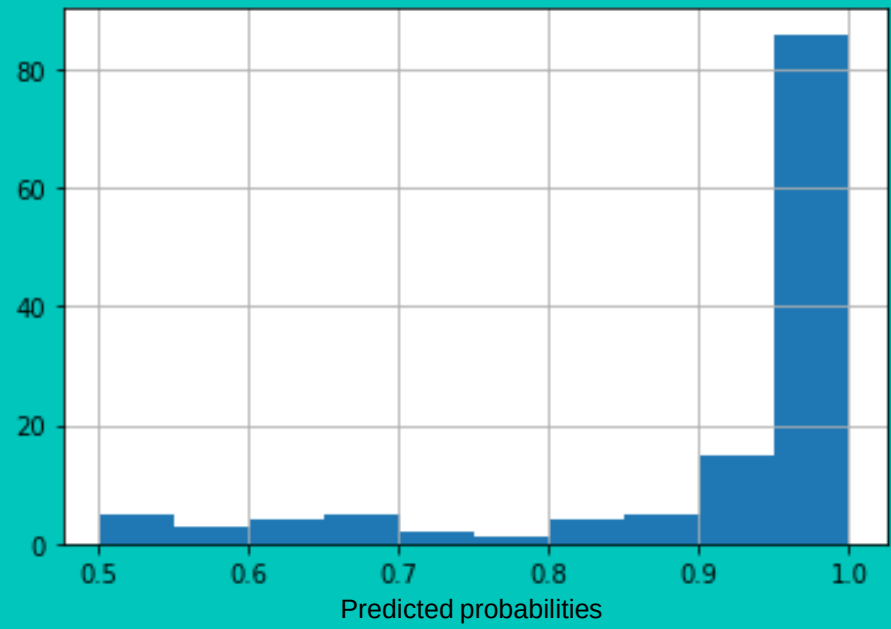
Second Layer Model Results – Period vs Modern

Second Layer Model Results – Histogram of Predicted Probabilities

Unseen images classified as Period (probability < 0.5)



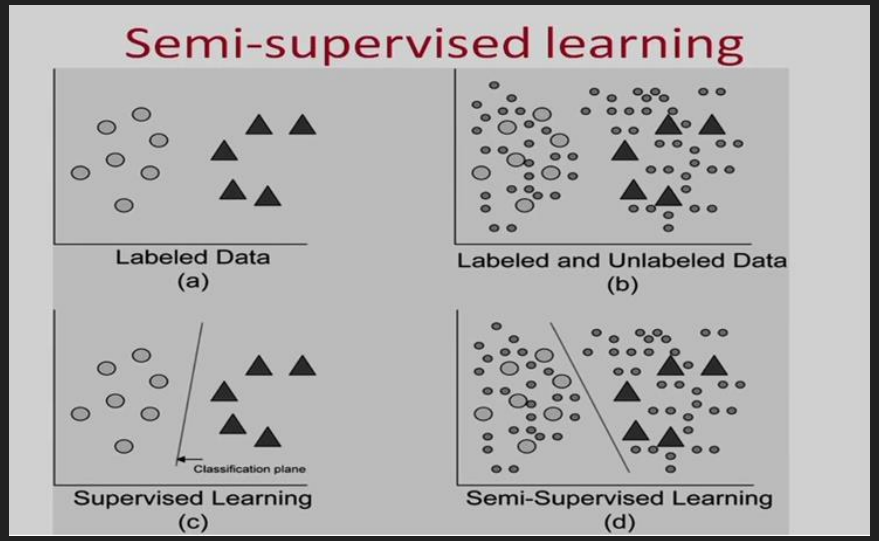
Unseen images classified as Modern (probability >= 0.5)



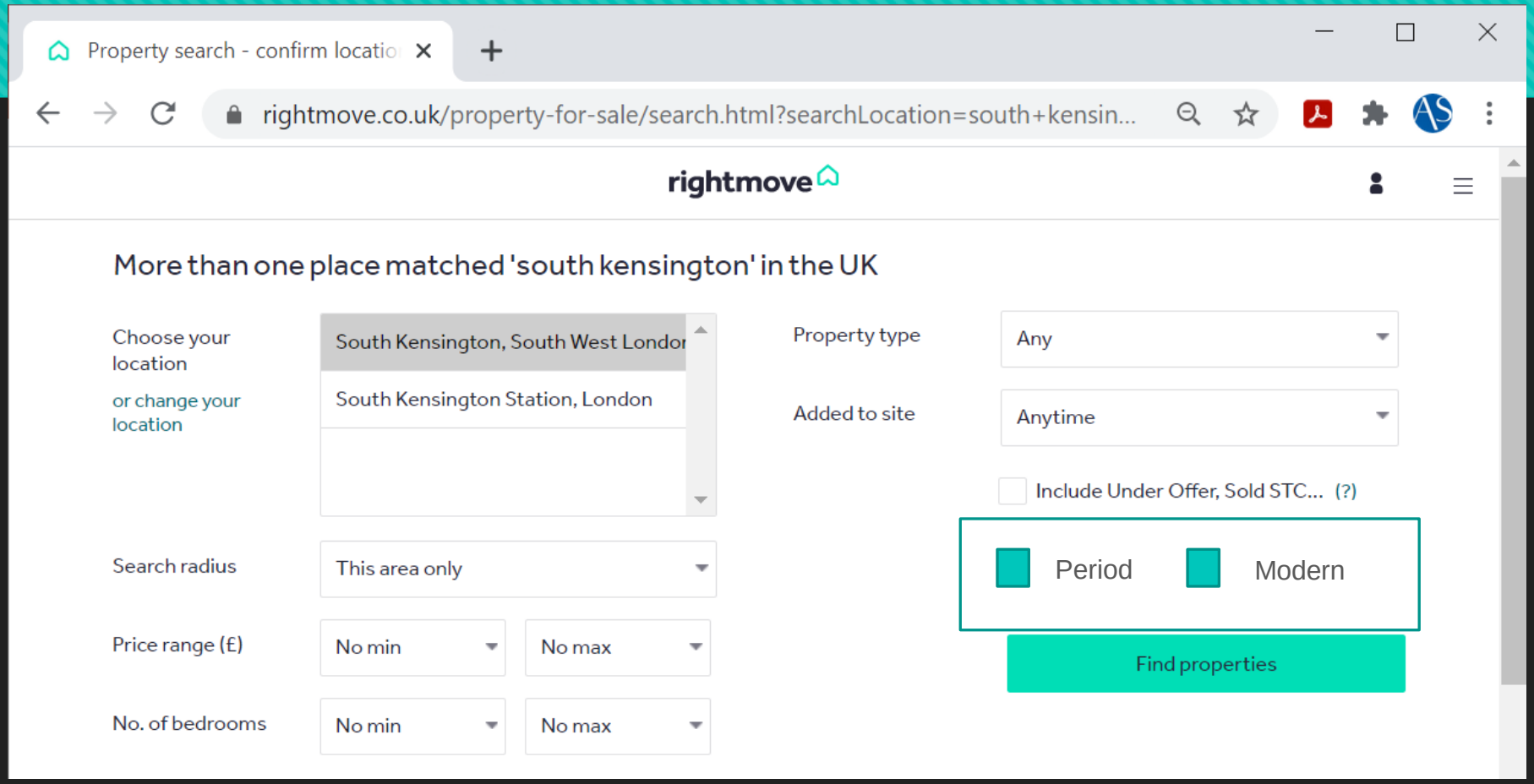
Majority of predicted probabilities are close to 0 and 1, ie high confidence of classification.
Confirmed by human eyes

MLOps Considerations

- Semi-Supervised Learning
  - Problem: update the trained model with new images
  - Restraint: limited labelled data
  - Proposal: prioritise the manual labelling of new images by looking at the misclassified cases at the extremes
- Further Feature Extraction
  - Exterior stye: Georgian, Victorian etc
  - Interior: shape of windows, burning stoves
- Aggregation of other data
  - Combine other data eg distance of landmarks, historical transaction price



Better Applications for Users



Conclusion

Conclusions

Property Tech

- A feasible and reliable process and modelling has been established to accurately classify period vs modern buildings
- With a divide-and-conquer approach, two layers of model were able to learn to classify two set of features
- Potential commercial applications in areas of property tech:
  - Property Search Engine
  - Property Recommender System
  - Property Investment Evaluator
- There is room for future development to fine tuning the model and aggregating the model outputs with other data currently available