USE OF MACHINE LEARNING FOR REAL ESTATE PROPERTY RECOMMANDATION

By : Patrick NSABIMANA

APPLIED DATA SCIENCE CAPSTONE PROJECT
IBM DATA SCIENCE PROFESSIONAL CERTIFICATE OFFERED THROUGH COURSERA

ABSTRACT

In capstone project , we observed how data science and machine learning approaches can be employed in the real estate industry. decisions to Buy properties are heavily influenced by the location of the houses. We showed how Python libraries such as Pandas can be used to statistically analyze the properties. we used data from FOURSQUARE to enrich the data with information on access to different facilities (venues) and used that to compare, score and rank the properties. The shortlist we arrived at can be used for field visits.

We conclude with a forward thinking approach to turn this into a recommendation model and suggest scope for future work in this area.

PROBLEM STATEMENT

With increased access to Internet through personal computers and mobile devices, many individuals turn to the Web to search for items of interest to them. However, the search for rental or buying properties can be a challenging task. Prospective tenants may spend a significant amount of time exploring various websites advertising properties for rent. Yuan, Lee, Kim and Kim (2013) found that searching real estate assets online does not benefit homebuyers in terms of time, flexibility and intuitive results.
Search engines generally do a good job by helping users find what they are looking for, however many people find it hard to match their preferences to a query that is likely to produce the desired search results, Viappiani and Faltings (2006) and Viappiani et al., (2006). Using recommender systems technology can address the problem of mapping user's preferences to items that are likely to meet them (Viappiani et. al, 2006).
A recommender system can be defined as any system that produces individualised recommendations as output or has the effect of guiding the user in a personalised way to interesting or useful objects in a large space of possible options, Burke (2002). Automated recommender systems emerged early 1990s. A good example is Amazon that uses the collaborative filtering approach in making recommendations on the basis of the current user's behaviour and that of other similar users, Ekstrand, Riedl and Konstan (2010). Netflix uses a combination of content-based filtering and collaborative filtering to make recommendations based on the current user's preferences and those of similar users.

Traditional approaches such as collaborative filtering and content-based filtering are not suitable for high-value items such as electronics, vehicles and real estate assets since they are not purchased as frequently as are other objects. Consequently, buyers are not able to leave a sufficient number of reviews

(ratings) on this objects to facilitate useful recommendation to other users, who may not be satisfied with years-old ratings, according to Felferning, Friedrich, Jannach and Zanker (2011). Knowledge-based recommender systems address this challenge by exploiting explicit user's requirements and the underlying product domain knowledge to generate recommendations, Felfering et al (2011).

METHOD and RESULTS

Data collection

Housing data has been extracted from DOF Condominium Comparable Rental Income from Brooklyn in the fiscal year 2011_2012, they came in a few CSV files of different sizes. Data was read using pandas as DataFrame objects. These DataFrames form the bedrock of both spatial and attribute analyses. The CSV files were merged to obtain an initial list of about 700 properties.

Data Cleaning

An initial and critical step in any data analysis and machine learning project is wrangling and cleaning of the data. In this case, the data suffers from duplicate sand illegal characters in column names,

We use Pandas to clean our data and come with a dataset of 626 rows to be used in our data analysis.

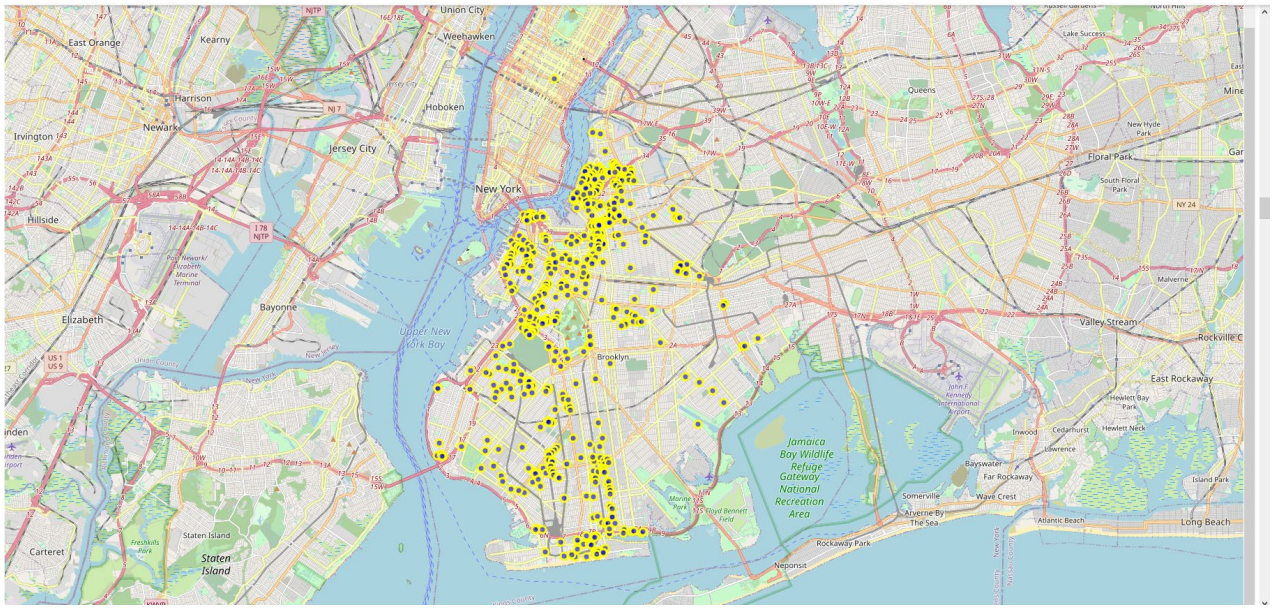
```
# read dataset hosted on our GITHUB repository
df=pd.read_csv('https://raw.githubusercontent.com/patrickns/coursera_captsone/main/dataset.csv')
df.head()
```

	city	bourgh	Neighborhood	Address	address_full	Building_Classification	Total_Units	Year_Built	Size	Value	Latitude	Longitude
0	New york	brooklyn	DOWNTOWN-FULTON FERRY	1 MAIN STREET	1 MAIN STREET,DOWNTOWN-FULTON FERRY,brooklyn,NY	R4-CONDOMINIUM	124	1913.0	227916	23759904	40.703629	-73.990410
1	New york	brooklyn	DOWNTOWN-FULTON FERRY	31 WASHINGTON STREET	31 WASHINGTON STREET,DOWNTOWN-FULTON FERRY,bro...	R4-CONDOMINIUM	13	2001.0	24672	2346000	40.703386	-73.989361
2	New york	brooklyn	DOWNTOWN-FULTON FERRY	133 WATER STREET	133 WATER STREET,DOWNTOWN-FULTON FERRY,brookly...	R4-CONDOMINIUM	69	2006.0	78352	8488003	40.703362	-73.989044
3	New york	brooklyn	DOWNTOWN-FULTON FERRY	50 BRIDGE STREET	50 BRIDGE STREET,DOWNTOWN-FULTON FERRY,brookly...	R4-CONDOMINIUM	58	1904.0	52530	4003003	40.703105	-73.984828
4	New york	brooklyn	DOWNTOWN-FULTON FERRY	4 WATER STREET	4 WATER STREET,DOWNTOWN-FULTON FERRY,brooklyn,NY	R4-CONDOMINIUM	13	2007.0	20085	2107998	40.703126	-73.993609

DATA ANALYSIS

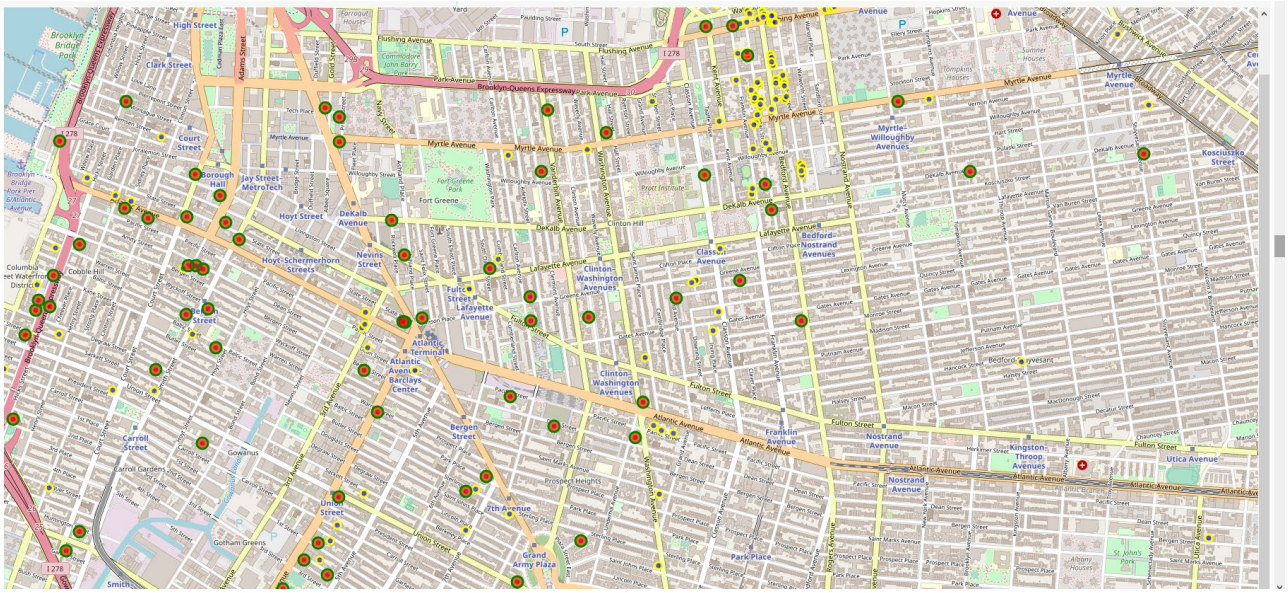
Exploratory analysis

We used pandas to Explore the distribution of numeric columns and used folium libraries to visually analyze our data



Running an Initial Shortlist

Basing on the houses' intrinsic features, we were able to build the first shortlist, the shortlist reduces the number of eligible properties from 626 to 258. When plotted on a map, these properties are spread across Brooklyn.



Quantifying Access

When buying a house, you are looking for proximity to services such as groceries, pharmacies, urgent care facilities, restaurant, spa and parks. The FOURSQUARE developer API was used to search for such facilities within a specified distance around a house.

```
[ ] brooklyn_grouped = brooklyn_onehot.groupby('Property').mean().reset_index()
    brooklyn_grouped
```

	Property	ATM	Accessories Store	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Austrian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Bask S
0	Prop001	0.0	0.0	0.0	0.033333	0.0	0.033333	0.0	0.0	0.0	0.0	0.033333	0.0	0.000000	0.0	0.0	0.0	0.000000	0.033333	0.000000	0.000000	0.000000	
1	Prop003	0.0	0.0	0.0	0.033333	0.0	0.033333	0.0	0.0	0.0	0.0	0.066667	0.0	0.000000	0.0	0.0	0.0	0.000000	0.066667	0.000000	0.000000	0.000000	
2	Prop004	0.0	0.0	0.0	0.066667	0.0	0.033333	0.0	0.0	0.0	0.0	0.066667	0.0	0.000000	0.0	0.0	0.0	0.033333	0.033333	0.000000	0.033333	0.000000	
3	Prop006	0.0	0.0	0.0	0.033333	0.0	0.033333	0.0	0.0	0.0	0.0	0.033333	0.0	0.000000	0.0	0.0	0.0	0.000000	0.033333	0.000000	0.000000	0.000000	
4	Prop007	0.0	0.0	0.0	0.033333	0.0	0.033333	0.0	0.0	0.0	0.0	0.033333	0.0	0.000000	0.0	0.0	0.0	0.000000	0.033333	0.000000	0.000000	0.000000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
253	Prop697	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.033333	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.033333	0.000000	
254	Prop699	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.052632	0.0	0.0	0.0	0.000000	0.052632	0.052632	0.052632	0.000000	
255	Prop700	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.038462	0.0	0.0	0.0	0.000000	0.038462	0.038462	0.038462	0.000000	
256	Prop702	0.0	0.0	0.0	0.047619	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.047619	0.000000	0.000000	0.000000	0.095238	
257	Prop703	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.040000	0.040000	0.000000	0.040000	0.000000	

258 rows × 303 columns

All the returned venues curated 302 unique categories

Data Scaling

While a scoring function can be extremely handy in comparing the features of shortlisted houses, when applied directly (without any scaling), it returns a set of scores that are heavily influenced by a small number of attributes that have numerically large values. For instance, an attribute such as property price tends to be a large number (hundreds of thousands) compared to the number of bedrooms (which will likely be less than 10). Without scaling, property price will dominate the score beyond its allotted weight. Scores computed without scaling appear extremely correlated with the property price variable. While property price is an important consideration for most buyers, it cannot be the only criteria that determines a property's rank.

To rectify this and compute a new set of scores, all numerical columns were scaled to a uniform range of 0–1 using the MinMaxScaler function from scikit-learn library.

	Value	Year_Built	Size	Total_Units	Gym	Bar	School	Park	American Restaurant	Spa	ATM	Playground
0	0.294919	0.119266	0.261681	0.195489	0.4	0.000000	0.0	0.175	0.1	0.0	0.0	0.2
1	0.099950	0.972477	0.082142	0.092105	0.4	0.000000	0.0	0.175	0.1	0.0	0.0	0.2
2	0.042692	0.036697	0.051145	0.071429	0.4	0.142857	0.0	0.350	0.2	0.0	0.0	0.0
3	0.194882	0.091743	0.177817	0.125940	0.2	0.000000	0.0	0.175	0.1	0.0	0.0	0.2
4	0.028560	0.091743	0.015877	0.001880	0.2	0.000000	0.0	0.175	0.1	0.0	0.0	0.2

Properties Scoring

Evaluating houses is a deeply personal process. Different buyers look for different characteristics in a house. Not all aspects are considered equally, so assigning different weights for features will let you arrive at a weighted sum (a score) for each house. The higher the score, the more desirable a house is to you. We used a scoring function that reflects the relative importance of each feature in a house. Desirable attributes are weighted positively, while undesirable attributes are weighted negatively.

	Property	Year_Built	Size	Value	Total_Units	Gym	Bar	School	Park	American Restaurant	Spa	ATM	Playground	scores_scaled
0	Prop001	1913.0	227916	23759904	124	0.066667	0.000000	0.0	0.033333	0.033333	0.0	0.0	0.033333	1.253935
1	Prop003	2006.0	78352	8488003	69	0.066667	0.000000	0.0	0.033333	0.033333	0.0	0.0	0.033333	2.491590
2	Prop004	1904.0	52530	4003003	58	0.066667	0.033333	0.0	0.066667	0.066667	0.0	0.0	0.000000	1.131438
3	Prop006	1910.0	158053	15924001	87	0.033333	0.000000	0.0	0.033333	0.033333	0.0	0.0	0.033333	0.974518
4	Prop007	1910.0	23150	2896003	21	0.033333	0.000000	0.0	0.033333	0.033333	0.0	0.0	0.033333	0.875971

Properties Ranking

Once the properties were scored, they were sorted in descending order and assigned a rank, creating a refined shortlist of homes that could be visited. In this example, the top 50 houses are spread across Brooklyn.

	Property	Year_Built	Size	Value	Total_Units	Gym	Bar	School	Park	American Restaurant	Spa	ATM	Playground	scores_scaled	rank
0	Prop508	1990.0	445921	11848942	552	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000	0.000000	4.047624	0
1	Prop571	2005.0	28750	2485000	21	0.000000	0.000000	0.000000	0.000000	0.000000	0.166667	0.000	0.166667	3.435403	1
2	Prop032	2008.0	628669	78988986	438	0.000000	0.000000	0.000000	0.100000	0.033333	0.000000	0.000	0.100000	3.395062	2
3	Prop515	2007.0	30000	3275999	23	0.000000	0.000000	0.045455	0.000000	0.000000	0.000000	0.000	0.045455	3.227646	3
4	Prop383	2006.0	39038	2998001	27	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.125	0.125000	3.218608	4
5	Prop372	2000.0	39000	3618992	32	0.083333	0.000000	0.000000	0.083333	0.000000	0.000000	0.000	0.083333	3.044450	5
6	Prop509	1990.0	425720	9955968	542	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	3.031429	6
7	Prop012	2006.0	301131	35211998	266	0.066667	0.033333	0.000000	0.066667	0.033333	0.000000	0.000	0.033333	2.922355	7
8	Prop642	1964.0	842968	21948992	371	0.033333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.033333	2.862696	8
9	Prop329	2001.0	85046	7759011	66	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.000	0.117647	2.642938	9

Building a Housing Recommendation Model

So far, the dataset was feature engineered with intrinsic and spatial attributes. Weights for different features were explicitly defined so properties could be scored and ranked. In reality, the decision-making process for buyers, although logical, is less calculated and a bit fuzzier. Buyers are likely to be content with certain shortcomings (e.g., fewer units) if they are highly impressed with some other characteristic (e.g., size). If buyers simply favor some houses and blacklist others, you could let a machine learning model infer their preferences.

Since it is difficult to collect this kind of training data for a large number of properties, a mock dataset was synthesized using the top 50 houses as the favorite group and the remaining as the blacklisted group.

This data was fed to a machine learning logistic regression model. As this model learns from the training data, it attempts to assign weights to each predictor variable (intrinsic and spatial features)

and predict whether that house will be preferred by a buyer. As a new property hits the market, this model can predict whether a buyer would like it and present only relevant results.

```
[ ]  coeff = log_model.coef_.round(5).tolist()[0]
     list(zip(X_train.columns, coeff))
```

```
[('Value', 0.86952),
 ('Year_Built', -0.80654),
 ('Size', 0.1225),
 ('Total_Units', 0.47617),
 ('Gym', 0.77884),
 ('Bar', 0.75015),
 ('School', -0.1432),
 ('Park', -0.0917),
 ('American Restaurant', 0.58796),
 ('Spa', 1.04237),
 ('ATM', -0.08805),
 ('Playground', 0.31364),
 ('scores', -0.14505)]
```

```
[ ]  log_model.intercept_
```

```
array([-1.27062641])
```

CONCLUSION

In capstone project , we observed how data science and machine learning approaches can be employed in the real estate industry. decisions to Buy properties are heavily influenced by the location of the houses. We showed how Python libraries such as Pandas can be used to statistically analyze the properties. we used data from FOURSQUARE to enrich the data with information on access to different facilities (venues) and used that to compare, score and rank the properties. The shortlist we arrived at can be used for field visits.

We conclude with a forward thinking approach to turn this into a recommendation model and suggest scope for future work in this area.