Rapportmall

08/10/2022

EC Utbildning Data Science

Gabriel Watson

Bakgrund, Syfte & Intressenter

Målet med detta projektarbete var att kunna få en inblick på AirBnb marknaden i USA. Det finns hundra tusentals av fastigheter runt om USA som läggs upp varje dag på Airbnb för att hyra ut till betalande kunder. För att få en bättre inblick på vart det bästa ställen att bo på i USA, i relation till pris, så kollar man på flera faktorer så som pris, läge, typ av fastighet samt vad fastigheten har att erbjuda en kund. I fallet på detta projekt så är det inte bara en kund som är en intressent utan även de som hyr ut på Airbnb för de kan själv få en inblick på de fastigheterna som är mest populära i landet.

För att kunna arbete med denna fråga så har vi hittat ett dataset ifrån kaggel som innehåller tusentals uthyrningar i USA. Några av features består av typ av fastigheter, pris och hur många som kan bo där exempelvis. Samt så finns det ett ratingsystem som ger inblick på de bostäderna med bäst betyg. För att avgränsa så kollar vi bara på några av de största städerna i USA

2. Hypotes
Mycket av det som väger om en fastighet är billig eller dyr att hyra ut ligger i vad fastigheten har för typ av bekvämligheter samt i vilket typ av område det ligger i stan. Exempel på detta kan vara ett stort hus i Los Angels kommer i snitt vara dyrare än ett mindre hus i samma stad. Samt att i en stad som Los Angels att fastighet med havsutsikt kan vara mer eftertraktat av kunder och därmed dyrare. Sen så kommer tillgängligheten spela en stor roll i snittet på de priser som man kan finna Airbnb fastigheter att hyra som en kund. Desto mer fastigheter om finns i en stad så kommer man hitta ett billigare genomsnitt medan de samtidigt kommer finns väldigt dyra samt väldigt billiga fastigheter.

3. Metod & Teknik-val
Projektet kommer att genomföras på ett agilt sätt vart laget kommer igenom verktyg som JIRA, att dela upp arbetet i mindre delare. Man arbetare på detta sätt för att på bästa möjliga sätt dela upp arbetet och få det maximala deltagandet ifrån varje medlem i laget. Sen så kommer man använda sig av Python för datahantering, data städning och prediktering samt som man kommer den använda sig av Power-BI för att visualisera det man har hittat i data igenom att skapa diskbords. Sen så kommer mycket av arbetet vara delat genom Github repositories så alla kan arbeta med koden och sen uppdatera den så alla har enkelt tillgång till det nya ändringarna.

4. Data & Källor
Vi hämtade vår Airbnb dataset ifrån Kaggle. Detta dataset innehåller tusentals av olika uthyrningar av fastigheter runt om de största städerna i USA. I detta dataset så finner man 28 kolumner och drygt 75 000 rader av data. Det innehåller flera olika features så som ID, log_price, proptery_type, amentiteis, bathrooms, city, latitude, longitude bara för att nämna några. Sen så fanns det även ett ratingsystem som ger inblick på de bostäderna med de bästa betygen. Många av de NA värden som fanns hittade vi i kolumnerna som vi inte behövde använda oss av i vårt projekt då de medförde inget för att besvara vår nyckelfråga. Därav kunde vi bedöma att kvalitén på datasetet var bra.

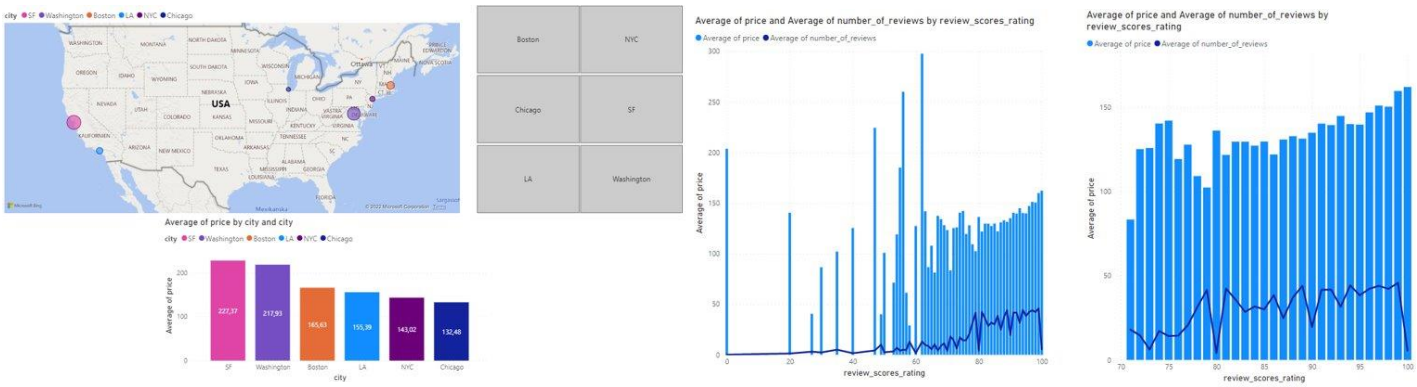
5. Genomförande

1. Det första som skedde var att hitta en databas som vi kunde arbete med och att ladda in de på visual studio code för att påbörja arbetet i python för att rengöra och utforska datan.

2. En av de första stora besluten som togs var vilka kolumner som vi kände att vi skulle ta bort och vilka som vi skulle ha kvar. De slutsatserna som vi drog var att om en kolumn inte hjälpte med att besvara vår nyckelfråga, så var den inte relevant att ha med. Ett exempel på detta var med kolumnen first_review dvs när den första omdömet gavs. Detta datum om när första omdömet gav är inte relevant för arbetet och därmed togs bort.

```
In [22]:
#Drop columns that is not neccesary for our model
airbnb.drop(["id","amenities","bed_type","description","first_review",
             "host_has_profile_pic","host_identity_verified","host_response_rate","host_since",
             "last_review","name","thumbnail_url","zipcode"],axis=1,inplace=True)
```

3. Efter vi hade städade och utforskat datan så förde vi över den städade datan till PowerBI där vi påbörjade de visualiseringar för arbetet. Här så var det mycket diskussion kring vilka typer av visualiseringar som vi ville ha. Eftersom vi hade hittade väldigt många key findings när vi arbetade med datan i python så ville vi även hitta ett par bra som kunde visualisera olika grejer.



4. När vi väll var nöjda med de visualiseringar som vi hade arbetat med i PowerBI så beslöt vi oss för att dela upp arbetet på presentationen och vilka delar som skulle presenteras av vilken gruppmedlem.

6. Key Findings & Main takeaway

- **Genomsnittliga priset för de olika städerna.** Kollade på det genomsnittliga priset för de olika städerna för att kunna avgöra vilka städer som var billigast respektive dyrast att bo I.
- **Korrelation mellan pris och rating.** Kollade på korrelationen mellan pris och omdömen för att se om antal omdömen påverkade priset eller om priset påverkade omdömen.
- **Prediktioner/observerade värden.** Vi har även försökt prediktera priset med hjälp av alla features som vi valt har varit signifikanta för oss. Vi valde att använda linearregression och XGboost.
- **Machine Learning.** Vi tittade lite extra på våra prediktioner där priset skilde sig väldigt mycket gentemot de predikterade värdena
- **Jämföra Europeisk stad med LA.** Här har vi valt att hämta ett airbnb dataset med uthyrningar från Barcelona. Vi ville jämföra en turist och kuststad i Europa jämfört med USA, så vi valde Barcelona att jämföra med LA

Den informationen som vi fick ifrån de följande keyfindings var:

- **Genomsnittliga priset för de olika städerna.** Gav inblick på vilka städer som va billigaste respektive dyraste. Detta i sin tur hjälper att ge denna inblick på marknaden i helhet. För beställaren så hjälper denna överblick bland annat med informationen vart fastighetsmarknaden kan vara dyrare och billigare i olika områden runt om USA.

- **Korrelation mellan pris och rating.** Gav inblick på om priset på en fastighet hade en korrelation på omdöme. Det man kom fram till var att ju högre pris desto mindre omdömen. Det är fler som hyr billigare bostäder och därav så ser man att det finns fler omdömen. Ju högre priser desto färre omdömen. Det kan vara att det är färre människor som har råd att hyra dyrt och därav mindre omdömen.
- **Prediktioner/observerade värden & Machine Learning.**
- **.** Vi använde oss av LinearRegression och XGboost för att prediktera priset med hjälp av alla de features som vi kände va viktiga för projektet. Det vi tittade lite extra på våra prediktioner där priset skilde sig väldigt mycket i genomförsels till de predikterade värdena och så att väldigt många av dessa hade inga omdömen. Sen så plottade vi även upp de observerade värdena på y-axeln de predikterade värdena på x-axeln. Majoriteten av värdena har predikterats ganska bra, dock så fanns det vissa undantag.
  Vi kan även se det genomsnittliga värdet på log_pris och vanligt pris mellan predikterade och observerade, vilket var väldigt bra.
- **Jämföra Europeisk stad med LA.** Denna punkt gav information i relevans till hur andra Airbnb marknader ser ut i andra länder. Vi ville jämföra en stad som är lik i Europa till en stad i USA, där av så valde vi LA och Barcelona. Då båda dessa städer ligger vid kusten samt att det är två städer med mycket turism och relativt likvärdiga städer i form av population och storlek. Det vi hittade var att LA var i snitt $45 dyrare än Barcelona, detta kan bero på flera olika anledningar men det vi såg var att alla städerna i USA var i snitt dyrare än Barcelona. Men det som var intressant att se hur mycket de skiljde sig mellan de olika städerna och om det fanns någon likhet mellan de olika Airbnb fastighetsmarknaderna.

Det man kan ta vidare ifrån dessa insikter är att om man hade haft mer data så som försäljnings data eller om man kolla på flera städer både i USA och i andra länder så kan man få ytligare mer information kring hur de olika priserna skiljer sig. Där av skulle mer data ge bättre och mer djupare inblick på de olika städernas priser och kunna ge mer information och nytta till beställaren.

Bifoga här gärna kod såväl som visualiseringar etc. (Räknas ej in i word- eller page-count)

Machine Learning

```
scaler = MinMaxScaler()
x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)

xg = xgb.XGBRegressor()
xg.fit(x_train_scaled,y_train)

print(mean_squared_error(y_pred=xg.predict(x_train_scaled), y_true=y_train))
print(mean_squared_error(y_pred=xg.predict(x_test_scaled),
                         y_true=y_test))
print(r2_score(y_pred=xg.predict(x_train_scaled), y_true=y_train))
print(r2_score(y_pred=xg.predict(x_test_scaled),
                         y_true=y_test))
```

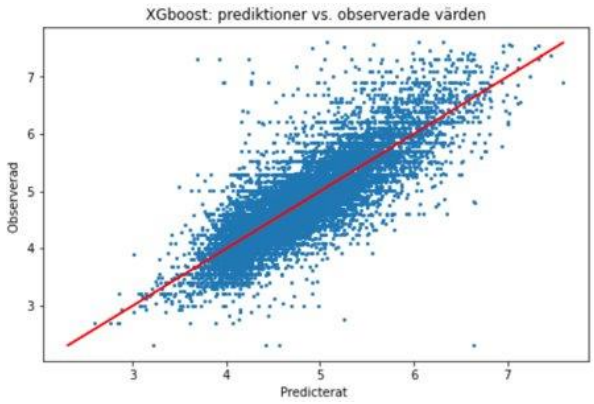
✓ 1m 16.9s

```
0.1274593976977311
0.1557412603288727
0.7514766113083144
0.7014721602768217
```

```
steps = [('scaler',MinMaxScaler()),
         ('MLR',LinearRegression())]
pipeline = Pipeline(steps)
params = {'MLR__fit_intercept': [True,False],
          'MLR__positive':[True,False]}
gridcv = GridSearchCV(pipeline,params,cv=3)
gridcv.fit(x_train,y_train)
print(gridcv.best_params_)
print(mean_squared_error(y_pred=gridcv.predict(x_train), y_true=y_train))
print(mean_squared_error(y_pred=gridcv.predict(x_test),
                         y_true=y_test))
print(r2_score(y_pred=gridcv.predict(x_train), y_true=y_train))
print(r2_score(y_pred=gridcv.predict(x_test),
                         y_true=y_test))
```

✓ 6m 5.5s

```
{'MLR__fit_intercept': False, 'MLR__positive': True}
0.1788406023972266
0.18792164383134685
0.6512923068346468
0.6397881829661795
```



	Actual	Predictions	int_price	pred_price
count	14823.000000	14823.000000	14823.000000	14823.000000
mean	4.789507	4.788140	162.788437	147.336517
std	0.722311	0.604649	174.632312	118.647926
min	2.302585	2.589319	10.000000	13.320697
25%	4.317488	4.317825	75.000000	75.025261
50%	4.718499	4.735275	112.000000	113.894753
75%	5.220356	5.177358	185.000000	177.213982
max	7.600402	7.589722	1999.000000	1977.762939

BARCELONA VS LA

18,84K
Total Listings

22,45K
Total Listings

110,27
Average of price

155,39
Average of price

Den viktigaste key findingen var det genomsnittliga priset för de olika städerna. Denna key finding var så viktigt efter som det var grunden kring majoriteten av de andra inblickarna som man fann samt att den hjälpte att besvara våran hypotes, kring att vi kunde prediktera priset på de olika fastigheterna. Denna informationen som vi fick ifrån de genomsnittliga priset var även viktigt för att kunna sedan jämföra de olika städerna runt om USA och att vi sen skulle kunna använda det som en bas att järnföra med städer utanför USA.

7. Slutsats

Det finns en korrelation mellan område och pris. Det verkar som att medans området spelar en stor roll på priset så finns det samtidigt andra faktorer så som omdöme och typ av bostad som har en betydelse för priset på en fastighet. Sen ser man även att priset på olika områden kan snare speglas i att det området överlag är billigare eller dyrare och att detta snare väger mer på priset än något annat. Sen det man skulle gärna vilja implementera skulle vara mer information kring mer bostäder samt att man skulle senare kolla på någon form av försäljnings data för att få en djupare inblick på de olika fastigheterna. Denna information skulle ge en direkt inblick på hur lönsam varje individuell fastighet var. Detta skulle i sin tur kunna förklara vilka typer av fastigheter är mest lönsamma att bedriva som Airbnb samt i vilka områden det är mest lönsamt att bedriva sin Airbnb verksamhet i. Sen så skulle man kunna få en väldigt bra inblick på de outliers som man finner i datan. Dvs om det finns några fastigheter som är inte lönsamma i extremt lönsamma områden och sen förstå sig på varför det kan

vara så. Sen även i områden som inte är lönsamma om det finns fastigheter som är extremt lönsam och försöka förstår på sig vad det är för typ av verksamhet de bedriva. Som en beställare så skulle detta vara väldigt intressant för då skulle de förstå vart och hur de skulle bästa göra en investering.