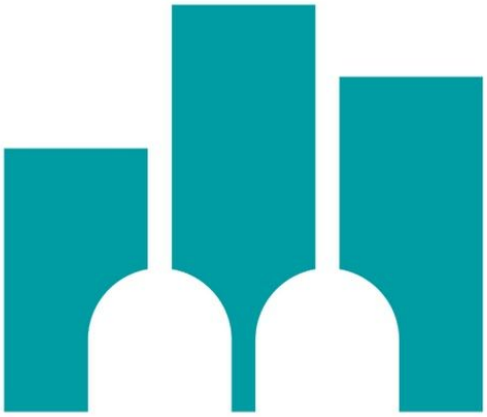


İSTANBUL
ŞEHİR
ÜNİVERSİTESİ

Graduate School of Natural Sciences - ECE

Big Data, Tools and Technologies

BAN 509

Instructor: Şaban Dalaman

Student Name: Taj Saleh
Student Number: 318183325

Subject: Project Report
Topic: Flights Recommendation System in Turkey using pyspark

Contents:

Introduction

The number of flights has increased dramatically in the past years, so did the data size that can be collected from various sources, with this popularity in the rise new services appeared to help customers around the world to check, choose and buy flights tickets to any destinations they want, Recommendation systems are one of those services that enable the customers to find flights with there own preferences such as: price, airlines, airline ratings, safety rating, duration of flight and least number of jets swapping to reach the desired destination, in this project i had the idea of introducing pyspark real time analysis as a part of a real time recommendation system for flights in turkey, the initial idea is very interesting and promising but new ideas has risen while doing this project such as Users Profiles and how can we sue them to better recommend flights options to users without them needing to manually choose it,.we went over collaborative filtering in our implementations with carefully designed mock users rating data for three airlines in turkey: THY, Pegasus, Anadolujet, we will talk in details about what we have done and what we can improve and add to it.

Data Collecting

Flights Data:

For Flights Data i used https://www.ucuzabilet.com/ to get flights informations which includes :

- Airline Name
- Flight Duration in mints
- Flight Departure Dates and Time
- Flight Arrival Dates and Time
- Flight Type (Economy, Business etc..)
- Departure City
- Arrival City
- Price
- Currency of price

This data will be generated when the user Choose his flight criteria such as date range (ex 3rd of june to 10th of june), we scraped data for the next 6 months but since flights prices change, this data should be scraped daily and checked for changes.
The code for this part can be found in ub-scraper.ipynb file.

Airlines Data:

For Airlines Data i crawled https://www.airlineratings.com to be able to get data for

- Turkish airlines
- Pegasus
- AnadoluJet

Data Scraped mainly included the following:

- Passengers Ratings
- Safety Rating
- Services and Product Rating
- Aircraft Type

This will be very useful in giving the user of the recommendation system extra criteria for choosing the flight, and not only be dependent on Time and Price.
The code for this is available in ar-scraper.ipynb file.

Users Ratings Data:

As mentioned in the introduction, i carefully worked on generating mock data for users rating for each airline, the Data includes 180000 data point, with 3 columns being:

- Airline id
- User id
- User Rating

since it will be used in Collaborative Filtering using pyspark MLlib, i had to make sure that this data is representative and i had to put some criterias that mainly can be listed in the following :

- Data should representative for each airline
- Data should contain enough data point for each airline
- Data should have similar number of data points for each airline

Ratings gathering assumed scenario:
3 airlines asks there customers to fill a rating paper after every flight, after a 2 years they end up with a large users rating dataset and make it public, now in this scenario the users rating data can tell you the following:

- The User liked the flight and continued his future flights with this airline
- The User did not like the flight and choose another airline company
- The User did not like any of the airlines and gave them bad ratings

In any case, this dataset can represent a profile for each user in terms of airlines and flights criteria.

Progress and Results

Data Analysis

Flights Data Analysis

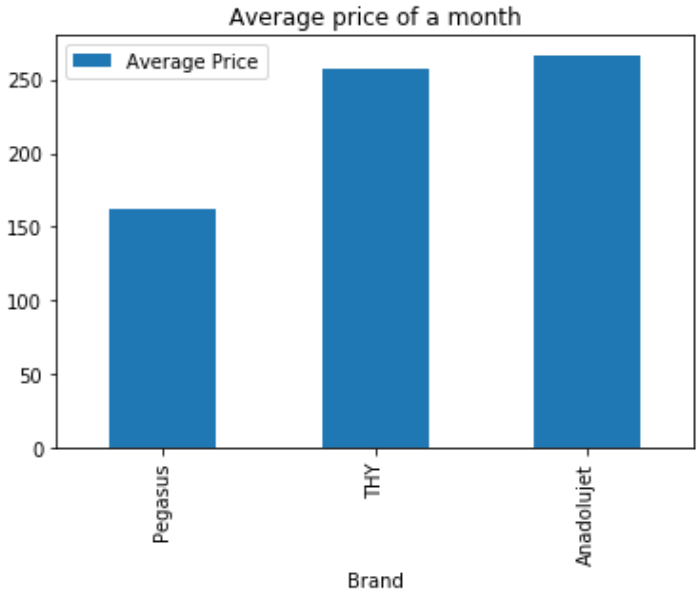
In this section I tried to make a basic flight recommendation based on Time and Price for flights from Istanbul to Ankara.

The Raw Data after few processing steps to use it as a pyspark dataframe:

Airline	Departure Time	Arrival Time	Flight Time(by mints)	Price	Currency	Type
Anadolujet	2020-06-02 14:30	2020-06-02 15:35	65	181.99	TRY	ECONOMY
THY	2020-06-02 10:00	2020-06-02 11:15	75	224.99	TRY	ECONOMY
THY	2020-06-02 14:10	2020-06-02 17:30	200	330.99	TRY	ECONOMY
Anadolujet	2020-06-02 12:00	2020-06-02 17:30	330	352.99	TRY	ECONOMY
THY	2020-06-02 11:15	2020-06-02 17:55	400	358.99	TRY	ECONOMY
Anadolujet	2020-06-02 13:30	2020-06-02 17:55	265	389.99	TRY	ECONOMY
Anadolujet	2020-06-02 10:40	2020-06-02 16:55	375	571.99	TRY	ECONOMY
THY	2020-06-02 13:00	2020-06-02 16:55	235	684.99	TRY	ECONOMY
Anadolujet	2020-06-03 14:30	2020-06-03 15:35	65	181.99	TRY	ECONOMY
THY	2020-06-03 10:00	2020-06-03 11:15	75	224.99	TRY	ECONOMY
THY	2020-06-03 14:10	2020-06-03 17:30	200	330.99	TRY	ECONOMY
THY	2020-06-03 11:15	2020-06-03 17:55	400	330.99	TRY	ECONOMY
Anadolujet	2020-06-03 12:00	2020-06-03 17:30	330	352.99	TRY	ECONOMY
Anadolujet	2020-06-03 13:30	2020-06-03 17:55	265	389.99	TRY	ECONOMY
Anadolujet	2020-06-03 10:40	2020-06-03 16:55	375	651.99	TRY	ECONOMY
THY	2020-06-03 13:00	2020-06-03 16:55	235	684.99	TRY	ECONOMY
Pegasus	2020-06-04 06:20	2020-06-04 07:25	65	162.99	TRY	ECONOMY
THY	2020-06-04 11:15	2020-06-04 12:30	75	193.99	TRY	ECONOMY
THY	2020-06-04 16:30	2020-06-04 17:45	75	193.99	TRY	ECONOMY
THY	2020-06-04 19:25	2020-06-04 20:40	75	193.99	TRY	ECONOMY

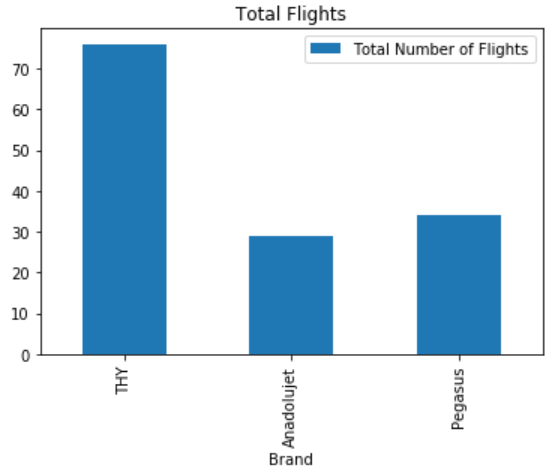
Average Price per airline :

Brand	Average Price
Pegasus	162.6958823529411
THY	257.7136842105262
Anadolujet	266.64517241379303



Total Number of Flights

Brand	Total Number of Flights
THY	76
Anadolujet	29
Pegasus	34



Average Price Per Airline for Each available Flight Time

Turkish Airlines

Brand	Departure Time	Average Price
THY	10	224.99
THY	11	242.7519047619047
THY	13	684.99
THY	14	330.99
THY	16	193.99
THY	19	272.1462499999999

Pegasus

Brand	Departure Time	Average Price
Pegasus	06	158.57823529411763
Pegasus	20	166.8135294117647

AnadoluJet

Brand	Departure Time	Average Price
Anadolujet	09	160.99
Anadolujet	10	222.40176470588236
Anadolujet	12	352.99
Anadolujet	13	389.99
Anadolujet	14	181.99
Anadolujet	19	444.99

Flights sorted by Price

Airline	Departure Time	Arrival Time	Flight Time(by mints)	Price	Currency	Type
Pegasus	2020-06-20 06:20	2020-06-20 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-12 06:20	2020-06-12 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-19 20:25	2020-06-19 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-15 20:25	2020-06-15 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-14 06:20	2020-06-14 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-18 06:20	2020-06-18 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-19 06:20	2020-06-19 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-06 06:20	2020-06-06 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-13 20:25	2020-06-13 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-16 06:20	2020-06-16 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-08 20:25	2020-06-08 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-09 06:20	2020-06-09 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-11 20:25	2020-06-11 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-09 20:25	2020-06-09 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-18 20:25	2020-06-18 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-10 06:20	2020-06-10 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-08 06:20	2020-06-08 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-10 20:25	2020-06-10 21:30	65	157.99	TRY	ECONOMY
Pegasus	2020-06-07 06:20	2020-06-07 07:25	65	157.99	TRY	ECONOMY
Pegasus	2020-06-11 06:20	2020-06-11 07:25	65	157.99	TRY	ECONOMY

From the past data analysis we can easily recommend the cheapest flight based on price and which time slot is better or cheaper for flights.
But what this part fails to do is give the customer more control over his options, a customer may be willing to pay extra for better services or higher safety, products and comfort ratings for the flight, similar data analysis based on those will be done in the next part.

Flights and Airlines combined Data Analysis

The Data after combining The airlines information dataframe with Flights data dataframe:

Airline	Departure Time	Arrival Time	Flight Time(by mints)	Price	Currency	Type	Safety Rating	Product Rating
Anadolujet	2020-06-02 14:30	2020-06-02 15:35	65	181.99	TRY	ECONOMY	10.0	7.5
THY	2020-06-02 10:00	2020-06-02 11:15	75	224.99	TRY	ECONOMY	10.0	8.5
THY	2020-06-02 14:10	2020-06-02 17:30	200	330.99	TRY	ECONOMY	10.0	8.5
Anadolujet	2020-06-02 12:00	2020-06-02 17:30	330	352.99	TRY	ECONOMY	10.0	7.5
THY	2020-06-02 11:15	2020-06-02 17:55	400	358.99	TRY	ECONOMY	10.0	8.5
Anadolujet	2020-06-02 13:30	2020-06-02 17:55	265	389.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-02 10:40	2020-06-02 16:55	375	571.99	TRY	ECONOMY	10.0	7.5
THY	2020-06-02 13:00	2020-06-02 16:55	235	684.99	TRY	ECONOMY	10.0	8.5
Anadolujet	2020-06-03 14:30	2020-06-03 15:35	65	181.99	TRY	ECONOMY	10.0	7.5
THY	2020-06-03 10:00	2020-06-03 11:15	75	224.99	TRY	ECONOMY	10.0	8.5
THY	2020-06-03 14:10	2020-06-03 17:30	200	330.99	TRY	ECONOMY	10.0	8.5
THY	2020-06-03 11:15	2020-06-03 17:55	400	330.99	TRY	ECONOMY	10.0	8.5
Anadolujet	2020-06-03 12:00	2020-06-03 17:30	330	352.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-03 13:30	2020-06-03 17:55	265	389.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-03 10:40	2020-06-03 16:55	375	651.99	TRY	ECONOMY	10.0	7.5
THY	2020-06-03 13:00	2020-06-03 16:55	235	684.99	TRY	ECONOMY	10.0	8.5
Pegasus	2020-06-04 06:20	2020-06-04 07:25	65	162.99	TRY	ECONOMY	8.5	6.0
THY	2020-06-04 11:15	2020-06-04 12:30	75	193.99	TRY	ECONOMY	10.0	8.5
THY	2020-06-04 16:30	2020-06-04 17:45	75	193.99	TRY	ECONOMY	10.0	8.5
THY	2020-06-04 19:25	2020-06-04 20:40	75	193.99	TRY	ECONOMY	10.0	8.5

Cheapest Price per Airline:

Airline	min(Price)
THY	193.99
Anadolujet	160.99
Pegasus	157.99

Highest Price per Airline:

Airline	max(Price)
THY	714.99
Anadolujet	651.99
Pegasus	212.99

flights for airlines with 10 safety score.

Airline	Departure Time	Arrival Time	Flight Time(by mints)	Price	Currency	Type	Safety Rating	Product Rating
Anadolujet	2020-06-09 10:00	2020-06-09 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-08 09:20	2020-06-08 10:25	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-07 10:00	2020-06-07 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-10 10:00	2020-06-10 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-11 10:00	2020-06-11 11:05	65	160.99	TRY	ECONOMY	10.0	7.5

lowest price of flights for airlines with 10 safety score

Airline	min(Price)
THY	193.99
Anadolujet	160.99

flights for airlines with at least 7 product and comfort score

Airline	Departure Time	Arrival Time	Flight Time(by mints)	Price	Currency	Type	Safety Rating	Product Rating
Anadolujet	2020-06-14 10:00	2020-06-14 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-19 10:00	2020-06-19 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-15 09:20	2020-06-15 10:25	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-13 10:00	2020-06-13 11:05	65	160.99	TRY	ECONOMY	10.0	7.5
Anadolujet	2020-06-16 10:00	2020-06-16 11:05	65	160.99	TRY	ECONOMY	10.0	7.5

Shows the lowest price of flights for airlines with at at least 7 product score

Airline	min(Price)
THY	193.99
Anadolujet	160.99

Other criterias exists on the notebook such as flights that have a promo, minimum flight time and Airlines min price per flight type.
The previous data explorations tells us how powerful this data can be when it comes to choosing a flight, so many variables and options and this should be made easy for the user to choose from, better yet, why not design a system that does it by itself for a registered user !!, we will talk about it in the next part.

Users Ratings Data Collaborative Filtering

We talked about how we generated Users Ratings Data, and explored simple data analysis and filtering methods using many criterias with spark sql dataframes, now it's time to explore pyspark MLLib and try its Collaborative Filtering techniques!!
We are going to build a recommendation model using ALS on training data after we split it.

Notes:
THY id is 1, AnadoluJet is 2, Pegasus is 3
Ratings range: 0 - 1
Number of users:15000 user

Sample Users Rating Data

airline_id	user_id	rating
1	13366	3
1	12693	3
1	10785	2
1	12387	2
3	12128	8
1	6206	2
3	11946	9
1	6616	3
1	1389	8
2	12391	2
1	8624	1
2	5242	10
2	3683	2
3	5806	2
3	11392	10
2	12289	1
2	8210	8
2	1093	0
1	8428	3
1	12362	1

Informations about the data:

summary	airline_id	user_id	rating
count	179976	179976	179976
mean	2.0	7499.499933324443	3.497605236253723
stddev	0.8164988492856836	4330.331511089922	3.1292329290142926
min	1	0	0
max	3	14999	10

We split our dataset into training set and testing set:

(training, test) = rawdataframe.randomSplit([0.8, 0.2])

Next we customized our parameters for ALS Model:

als = ALS(maxIter=10, regParam=0.02, userCol="user_id", itemCol="airline_id",ratingCol="rating",coldStartStrategy="drop",nonnegative=True)

After successful execution of spark jobs its time to evaluate the build model using inbuilt transform function. This function is more or less similar to predict() function in the traditional machine learning algorithm(Sklearn). However, transform () function transforms the input test data or unseen data in order to generate predictions.

predictions = model.transform(test)
predictions.show()

After we transformed our model with the testing set, we printed out the first 20 rows of predictions and the results seems very close to their real values.

airline_id	user_id	rating	prediction
1	148	5	6.7196846
1	471	10	6.9970956
1	496	7	5.711477
1	1088	5	6.341976
1	1645	8	7.646962
1	2122	5	7.9662423
1	2366	10	7.2921357
1	3749	8	6.7230973
1	3794	8	7.648612
1	3918	7	6.086511
1	3918	10	6.086511
1	3997	5	7.292645
1	4101	7	7.967496
1	4519	5	6.3930793
1	4818	10	7.013397
1	4900	7	8.245024
1	4935	9	7.65016
1	5156	0	0.6502005
1	5518	0	1.4087081
1	5518	2	1.4087081

More exploration, Try to check all prediction for one user in the testing Data:

```
recomendations = model.transform(single_user)
recomendations.orderBy('prediction',ascending=False).show()
```

```
+----------+-------+------+----------+
|airline_id|user_id|rating|prediction|
+----------+-------+------+----------+
|         1|    471|    10| 6.9970956|
|         3|    471|     2|0.34230512|
|         3|    471|     3|0.34230512|
+----------+-------+------+----------+
```

We understand the following from this output:

- The user liked His flight with THY and gave it a high rating.
- The user disliked Pegasus and gave it a low rating.
- The user tried Pegasus more than once and rated it more than once, and in both cases he didn't like it.
- Based on this output the System would recommend Turkish Airlines for this User.

This seems well and good since the Model was able to detect the preference of a user, based on this output when the user tries to search for flights in future, the system understands his preference for THY airlines and proceeds to give him other criteria such as the ones we talked about in previous sections.

Conclusion and Future additions

In Conclusion, we understand that Flights Data when combined with Airlines ratings and informations can be very powerful in any flights recommendations systems, we also explored the use of Collaborative Filtering for the purpose of recommending Airlines for a specific users, this proved to be very powerful in understanding the users preference giving that we have enough data on the user, an addition to this model can be is the addition of other metadata such as price and time data exploration to the Model to better understand the user preferences and i expect that this will hold a much powerful results for the recommendations system and accuracy on users preferences .