Udacity Capstone Spark Project

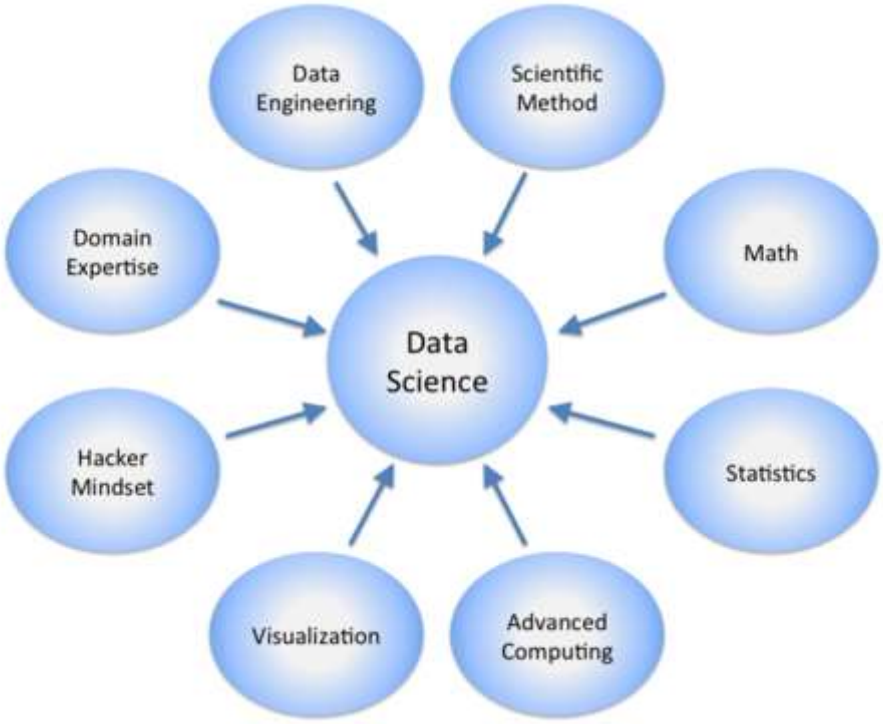
Rodrigo Santana Follow
Sep 18 • 4 min read

. . .

Hi All,

My idea with this post is to quickly show my final impressions at my Data Science learning Path as well as my 1st Spark project

A lot of learning and to learn



There was a very good experience at the Udacity nanodegree program with a lot of different topics/disciplines and a lot of learning. Interesting to see how much broader the subjects can get if you look for references. I can see the Data Science topic as a never-ending studying discipline, which for me is very cool :)

The Sparkify Project

In this problem based learning project, we are using a tiny database with some user logs from an application called “Sparkify” with emulates an music application like Spotify.

As normal music apps, it has free and subscribed users and our intention is to study user actions at music app logs in order to make sure they keep subscribed(user churn). The called user churn is one of the most impactful risk from an application loosing revenue .

An additional requirement was to use Spark library to handle in memory data manipulation / analysis.

My main steps were the following ones:

- Load and Clean dataset
- Data exploration & analysis
- Feature engineering
- Modeling
- Testing
- Conclusions

Load, exploration and data Analysis

Load data is very similar to Pandas so no news in this area.

Then I started eploring the information we have in the data. It was nice to see how easy it is to use SQL statements with spark:

```
spark.sql("SELECT DISTINCT(page) \
            FROM log_table\
            ").show(truncate=False)
```

```
+----------------------------+
|page                        |
+----------------------------+
|Cancel                      |
|Submit Downgrade            |
|Thumbs Down                 |
|Home                        |
|Downgrade                   |
|Roll Advert                 |
|Logout                      |
|Save Settings               |
|Cancellation Confirmation   |
|About                       |
|Settings                    |
|Add to Playlist             |
|Add Friend                  |
|NextSong                    |
|Thumbs Up                   |
|Help                        |
|Upgrade                     |
|Error                       |
|Submit Upgrade              |
+----------------------------+
```

One hint though is to remember to create a memory table before we could execute sql statements using "createOrReplaceTempView" function.

Feature engineering and data clean

Once exploration is done, we start some data cleaning and feature engineering as we know the data never is "ready for use" in any models we want to apply. In this case, couple of transformations and data reports were created in order to prepare it for model application. Examples:

- Transform timestamp at "normal time" DD-MM-YY
- Same as above but for date/time user registered
- Get for how long user was using app

- how many users are subscribed vs free
- how many have added Thumbs up or Thumbs Down
- how many songs have they played and how many artists have they listened to
- how may adds have they received

	level	TU	TD	songs	artists	usage	advert	churn
0	0	17.0	5.0	269	252	11503.553241	1	0
1	0	21.0	6.0	378	339	70074.629630	1	0
2	1	21.0	6.0	378	339	70074.629630	1	0
3	0	0.0	0.0	8	8	71316.886574	1	1
4	1	100.0	21.0	1854	1385	10619.814815	0	1

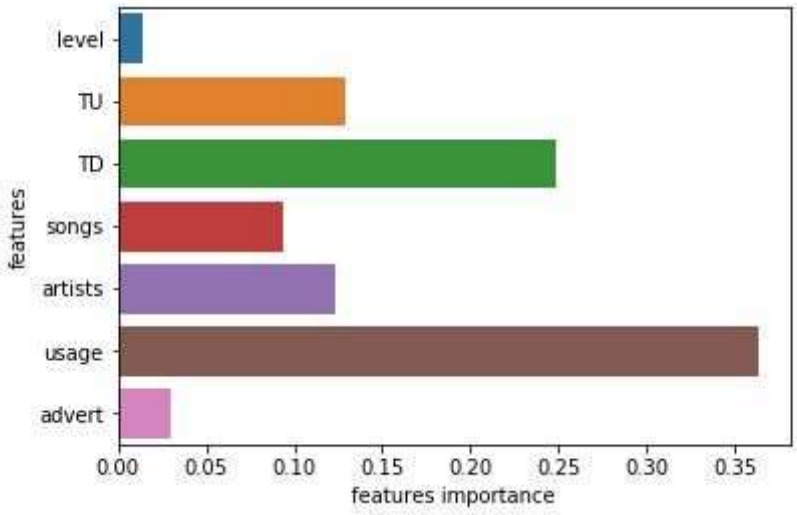
Sample of the final data

Modeling and Testing

I have split the data into train and test in order to be able to run the models at one sample data and test it in another one.

Once we got all this info I was able to apply it into 2 models, Logistic Regression and Random Forest as this problem is a classification type “possible leavers or not”

Metrics & Results



Since the churned users are a fairly small subset, I used F1 score as the metric to optimize.

I could get 0.91 (f1 score) with Random Forest

I was able to see that as much as the application is been used and the Thumbs Down is select for it more likely users will unsubscribe from the app.

So a good counter measure will be giving discounts or free month subscriptions for long time users that started disliking the service.

Conclusion

In an overall conclusion, the studies at Udacity nanodegree were a lot fruitful in order to better understand how broad a Data Scientist job can be and also add for me more passion for this subject :D

All the projects were adding a lot of value during the degree and spark one was a great addition to it with different concepts and application.

This project was good opportunity to play with spark library, handled things in memory.

We started with some “raw” data, loading it with spark, cleaning it and transforming some features in order to make the data more “ready to use” Once it was done we could split the data in train and test, create some pipelines in order to run 2 different classification models with best one with 0.83 F1 score.

We can see that running random forest we could get better f1 score than using Logistic Regression and it seems the feature with most impact is how usage time and thumbs down. so a good trigger for possible cancellation is the raise of Thumbs down from the user, so we can send discounts or something similar.

Udacity Spark