Big Data Tools 2 – 2021-2022: Group Project

Prof. Dr. S. Hoornaert

Context:

The Covid-19 pandemic and the first lockdown in March-May 2020 has created a significant negative impact on the world economy. One industry that was hit very hard, is the leisure industry including recreation, restaurants, bars, and other social gathering hotspots. These businesses are looking for innovative ways on how they can reach their customers and are exploring how online tools could be used for this purpose. One of these tools is Yelp (yelp.com), a social media style customer review website where customers can post their experiences with businesses.

Assignment:

Your team is hired by Yelp to identify and predict what factors lead some businesses to start doing delivery or takeout for the first time after the first lockdown. Yelp has provided two datasets from two different periods: the JSON dataset is from March 2020 while the Covid-19 dataset is from June 2020 (see the section on "Data" below). Your job is to build a prediction model to predict which businesses will start doing delivery/takeout after the first lockdown (use the variable "delivery or takeout" in the Covid-19 dataset for this purpose). Your goal is to build a very good prediction model (i.e., high AUC) using the data provided. You can be creative and innovative on how you use the available information (e.g., create new variables, use unstructured content, etc); as you would do in practice! The team that achieves the highest AUC on a 10-fold cross-validation approach gets +1 bonus point on their assignment; respecting of course the appropriate modeling setup process (e.g., no AUC-hacking or other methods of cheating)! Describe your approach in the technical section of your report. This section should be concise and written for a data science audience (e.g., describe creation of variables, algorithms used, cross-validation approach, evaluation metric(s)).

In addition to focusing on prediction, also provide insights on which criteria are important for determining whether a business will start doing delivery or takeout. Think of 2 creative ways (e.g., website features, sending a direct email) on how Yelp could target the businesses that fit these criteria (e.g., business profile) and that are already active on their platform. Yelp knows from previous reports that it has a 33% higher likelihood of selling ads to businesses that are more advanced in their digital transformation. Describe these elements in the business section of your report. This section should be written for middle to senior management that are responsible for business development.

Project organization:

This group project is organized in groups of 3 people. The groups were created at random. You can find the groups and their composition of team members on ieseg-online.

Delivarables:

By **Sunday, April 3 (23:59)** upload the following items under the assignment dropbox 'Group Project' on ieseg-online:

- A high-quality **report** with 1/ an executive management summary (1 page) in which you summarize the project's setup and main conclusions + 2/ a business section in which you discuss the relevant insights and business actions for Yelp (2 pages, incl. figures/tables) + 3/ a technical section (2 pages, incl. figures/tables) in which you explain the methodology (e.g., definition of variables, variables included in the model, algorithms used, cross-validation technique) and your reasoning in more detail. You can use academic articles or other secondary materials in your report to motivate your methodological approach and findings. Think about visualization and write in a concise way.

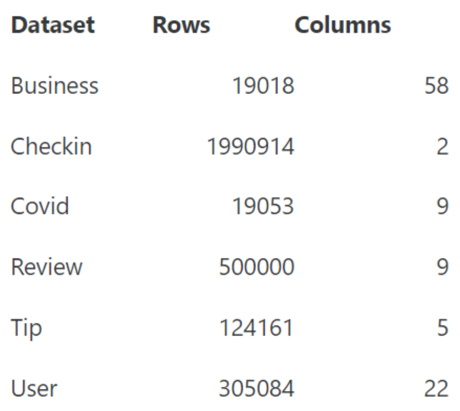
- A **notebook** created in Databricks using Spark with all code to execute the assignment. For this notebook:
  - In the first cell, write your team members' names, the academic year, and the course name. In the second cell, provide a "path" variable so that I can set the path of the file names on my own account and run your code.
  - Write fast, efficient and well-structured Spark code statements. Use Spark's functionalities as much as possible (e.g., pipeline(s), model building, tuning)!
  - Act as a professional Big Data scientist and document your code well. Make sure to stick to the checklist for coding best practices on https://www.topcoder.com/coding-best-practices/.
  - As a reminder, you can download your notebook from Databricks using "File" -> "Export" -> "Source File". Name this .py file in the following format: "BDT2_2022_NameTeamMember1_NameTeamMember2_NameTeamMember3.py"

Requirements:

- Use at least two algorithms in your modeling phase.
- Use at least two performance metrics; one of which should be the AUC.

Data:

You can download the datasets from ieseg-online. You can also find a description of the different variables per dataset there. In the table below you can find the dimensions of the datasets.



Dataset	Rows	Columns
Business	19018	58
Checkin	1990914	2
Covid	19053	9
Review	500000	9
Tip	124161	5
User	305084	22

The concept of this assignment is that you learn how to work on a Big Data problem independently. By solving problems yourself, you learn how to overcome them and improve your learning of working with Spark significantly. In case of problems, you can send me an e-mail (s.hoornaert@ieseg.fr).