royjafari / optimizing-big-data-problem-statement Public

Code Issues Pull requests Actions Projects Wiki Security Insights Se

main

optimizing-big-data-problem-statement / ch3 / Solutions.MD

royjafari Ch3_NewSolutionAdded History

1 contributor

142 lines (84 sloc) 14 KB

# Answers to Case Studies - Chapter 3

## Study tips

- Make sure to have given the questions a try before reading my answers
- The provided answers are one of many possible and correct answers that can be given to the raised questions.
- It will be great to compare your answers with my answers and use common sense and logic to find the better parts of each answer.

## Case Study 2: Point of Interest (POI) Recommendation

**Q1. Show your Business Understanding using the five questions provided in this chapter 2.**

To make sure that we have a proper business understanding, we will evaluate our ability to provide reasonable answers to the following questions.

**Who owns the risk and benefits? In other words, who is the decision-maker?** Yelp.com.

**What is the Decision?** What POI to recommend.

**What is the challenge of decision-making?** People are very different in taste and behavior.

**What are the risks and benefits?** Yelp.com may lose money if the POI recommendation does not add to the app user's engagement with the platform. On the other hand, Yelp.com will benefit if the POI recommendation leads to an engagement increase.

**How can data analytics create value?** The historical behavior pattern of the users may give us a sense of the app users' uniqueness, and the collective data of all the users may give us general patterns of interest that can lead to the creation of successful POI.

**Q2. Mention at least one possible source of data that can be useful for this case study & provide and provide an answer to the nine questions provided under Data Understanding in Chapter 2.**

**How is the source of data collected?** App usage data

**What are the Volume, Variety, and Velocity levels of each source of data?** The variety of the data is low, but the volume and velocity will be very high. Every tap on the screen will be one data point, with over 178 million unique visitors, with an average usage of weekly visits, and 3 screen taps per visit, we should expect about 300 new data points per second.

**What are the assumptions made in recording each value in each source of data?** 1) the recorded screen taps are intentional (not accidental), 2) the recorded taps come from real humans and bots are not generating fake data.

**What are the legal and ethical implications of using each source of data?** For instance, GDPR protects European citizens, if Yelp.com does not have their permission to use their usage data, we cannot use those records.

**What kind of values do we normally get in each column in every source of data?** Each data object will have the user's unique id, the IP address used to connect to the server, and the time and location of the user.

**Where is the data source in the spectrum of raw to fully processed data?** Access to the rawest possible data is not unimaginable.

**How the source of data could be useful?** The source of data could be used to extract unique patterns for each visitor, and also find collective patterns among all visitors.

**Have you been given access to meaningful samples of the data sources, and spent time getting to know them?** This has to be answered for real situations.

**Q3. Use the Business and Data Understanding from Q1 and Q2, to come up with a Viable Problem Statement.**

*‘Can we find generalizable patterns in the historical usage data that lead to the generation of meaningful POI recommendations?’*

**Q4. Between supervised, and unsupervised modeling, which type could make solve the viable problem statement you came up with?**

Unsupervised methods such as Clustering or Association Rules Analysis will have to be used because it is impossible or too expensive in this case study to have dependent attributes.

**Q5. Design a dataset that could be created from the data sources you identified under Q2. An example of such design can be seen under Case Study 1: Employee Relationship Management, Stage 2 Data Acquisition & Integration.**

Based on the method we decide to use for this case study the designed dataset will look very different. The following table design is used for clustering analysis.

**Data Object Definition**: one week of

| Column | Description |
|---|---|
| Gender | The visitor’s gender if available |
| Age | The visitor’s age if available |
| City | The visitor’s city of residence if available |
| State | The visitor’s state of residence if available |
| Country | The visitor’s country of residence if available |
| CityMode | The most common city where the visited POI pages are located |
| RestaurantPercentage | The percentages of restaurants among the POI pages visited |
| MusiumPercentage | The percentages of museums among the POI pages visited |
| MarketPercentage | The percentages of markets among the POI pages visited |
| NaturePercentage | The percentages of nature POI among the POI pages visited |

**Q6. Mention at least one useful Data Transformation and Feature Extraction that could help solve the viable business problem.**

**Data Transformation**: the following four attributes are the results of data transformation: `RestaurantPercentage` , `MusiumPercentage` , `MarketPercentage` and `NaturePercentage` . The raw data is a list of POIs that the user has shown interest in, and we transform that data by calculating these four attributes for the list.

**Feature Extraction**: the attribute `CityMode` is the result of feature extraction The raw data will be a list of GPS locations that the users have been known to be at, and based on that we will first transform the data into the list of the cities that the user has been at, then use the mode of the list as the value for `CityMode` .

**Q7. How can the model that you will create using the prepared data be validated? What metrics should be used?**

There can be two types of evaluations: Before using the model & after using the model.

Before using the model: Clustering analysis does not have a dependent attribute so we cannot use metrics such as accuracy, precision, or recall. We can however calculate some values to measure the quality of clustering such as the Sillohoute Score. There is one caveat about such metrics that do not rely on dependent attributes. These metrics can only be used to compare methods with one another and their value itself does not tell us much about how well the clustering has been in absolute terms. So in this case study, we will only be able to use the s Sillohoute Score to realize which clustering algorithm will give us qualitatively better results.

After using the model: After we use the insights in the clustering to suggest POIs to the users we can measure their engagement with the suggested IPOs and compare them with the rest of the links on the app. Furthermore, we can do A/B testing to compare the overall engagement of users on the platform for the people who the model is used for suggesting and for the people that are not suggested POIs.

**Q8. Mention one example of the difference between the following items.**

- **Data Acquisition in stage 2 vs Ingestion in Stage 5**
- **Data Transformation in stage 3 vs Transformation in Stage 5**
- **Modeling in stage 4 vs Modeling in Stage 5**
- **Validation in stage 4 vs Validation in Stage 5**

**Data Acquisition in stage 2 vs Ingestion in Stage 5**: For instance, in stage 2 we may have to be on an extensive email exchange with the owner of the GPS database to get GPS data of sample users over an extended period, however, in stage 5 the GPS data must be streamed into the application endpoint so the latest data can be used for creating recommendations for the user.

**Data Transformation in stage 3 vs Transformation in Stage 5**: in stage 3 all the data transformations are experimental so for instance, if we use python for loop to find the frequencies of all the cities before arriving at the CityMode, that’s fine, however in Stage 5 we want to have optimized code so using such unoptimized coding will not be accepted as much or it will be too costly.

**Modeling in stage 4 vs Modeling in Stage 5**: in Stage 4 modeling includes playing with clustering algorithms and also their hyperparameters to find the best algorithm/slover. However, in stage 5 modeling will simply repeat the training of the best algorithm/solver with the new data.

**Validation in stage 4 vs Validation in Stage 5**: Since we are using an unsupervised learning paradigm, the validation can only be intrinsic in both stage 4 and stage 5. Please read the answer to question 7. The difference between stages 4 and 5 is that in stage 4 we can take our time to experiment with all the possible algorithms and hyperparameters, but in stage 5 we can only have a handful of algorithms and hyperparameters.

**Q9. What are the metrics for Model Selection in stage 5, and how those metrics should be different from the ones in Model Validation under both stage 4 and stage 5?**

The metrics will be the same because we don’t have a lot of different metrics for unsupervised learning. However, in stage 5 we may decide not to update the trained model after all and go with the one that we already have in place.

## Case Study 3: Dynamic Pricing

# Feasibility Analysis

## Business Understanding

This document provides information to solve the following problem.

*“Is it possible to meaningfully and profitably develop an hourly dynamic pricing, and if yes, are the benefits of hourly dynamic pricing outweigh the costs of developing such capability?”*

## Data Understanding

The listed current sources of data:

- **DVI data**: Monthly unique visitors sent to us on the last Friday of every month from Data Vision Incorporated (DVI)
- **Webscraped data**: weekly prices scraped from competition’s websites

To be able to have hourly dynamic pricing we need the velocity and aggregation level of our data sources to be at least hourly or faster than hourly. The following list of our findings from these two sources of data.

**DVI data**: we currently pay $399 to DVI for the monthly data feed. DVI does not provide hourly data feed services and the fasted service they offer is daily, and they will charge $3999 for the daily data feed. One of the DVI competitions provides the hourly data for a $16,000 yearly bill. The same competition charges $5000 for the daily feed. We have to pay attention that if we decide to switch our data feed from DVI to Home Run Data (HRD), the competitor, it will take one of our software developers about 3 months to update all our ingesting solutions so they would match with HRD; this switch will have a one-time extra expense of $21,000.

**Webscraped data**: This data is collected, maintained, and made available to the business by an internal team. After communication with the team in question, they indicated they can perform hourly data scraping with minimal development costs, however, they estimated that our bill from Amazon Web Services (AWS) will jump from about $120 a month to $2200 a month due to the extra computation and data bases usages. During the call, one of the team members indicated that in the current scraped data they have seen that most of the competitors only update their prices every quarter and there is only one competitor that updates their prices every week; so to stage ahead of all the known competition they fastest scraping data that we will need will be perhaps daily, and for that, we will only increase our AWS bill to about $400 a month.

## Data Preparation

The data preparation will not change and we only need to perform whatever we did more often.
Modeling:

This project does not involve any new modeling, we just need to perform the optimization mode more often.

# Deployment

The deployment will be similar to our current system however we will need to run our models more often. For hourly pricing, 720 (24*30) more times; for daily pricing 30 more times.

Currently, our monthly bill from Google Cloud Platform (GCP) is $80 per month for the current speed of modeling. The estimated monthly bill for hourly deployment will be about $27,000 and for daily deployment will be $2,000. Since the same models will be run just more frequently there will be no initial cost of development.

## Summary of Options and their Costs

**Daily Pricing**: the total initial cost for daily Pricing will be zero. We will have to pay $3600 (3999-399) more to DVI yearly and our AWS bill will increase by $4,800 (12*400) yearly; moreover, our GCP bill will also increase by $21,600 yearly – (2,000-80)*12. Our yearly cost of pricing will increase by $30,000 per year.

**Hourly Pricing**: The total initial cost for this pricing will be $21,000. Furthermore, we will have to pay $15,601 (16000 - 399) more money to get hourly page visit feeds. The cost of web scraping will be the same as daily pricing (a yearly increase of $4,800) as we came to the conclusion the fastest web scraping that can add meaningful value is daily. However, the deployment cost will increase by $323,040 – (27,000-80)*12. In total, we will incur 21,00 initial cost, and the yearly cost of pricing will increase by $343,441 (323,040+4,800+15,601). For the sake of comparison, we will use convert the initial $21,000 into a 10-year annuity using a 0.15 interest rate. The 10 years comes from the fact that we use new pricing systems on average for 10 years before we go to a newer better system, and the 0.15 comes from the fact that the rate of return of the retailer is %15. The annuity of $21,000 with the mentioned parameters will be $4,184.29. So in total, we can estimate that the hourly pricing will increase our cost by $347,625.29 (343441 + 4184.29) yearly.

# Conclusions

On average we sell 20 million items every year. The average cost of our items is $88, and the average gross margin we get from our products is $15.5. For the daily pricing to make sense, if the new pricing system can increase the average margin by $0.0015 (30,000/2,000,000) or more the new daily system will be profitable. The breakeven point for the average making in the case of hourly pricing is $0.17381.

This feasibility analysis does not answer the question of whether either pricing can lead to such increases in gross profit. For answering those questions we will need to perform pilot studies and measure the success of those pricing systems in increasing the average profit margin.