HW11 First steps of the project

Car rental demand detection based on price requests from online channels.

Pirge Kaasik, Joonas Puura

Link to presentation slides:

https://docs.google.com/presentation/d/1veA_WQcfRRx7hQnE8qklmsLYceWzQGrPHrieSEWqcaI/edit#slide=id.g2a549964dd_46_0

EX1. Business understanding

Identifying your business goals

RateChain is a company that provides price management solution in online channels for car rental companies all over the world. RateChain calculates price quote for each request from online distributor. Since there are about ca 2 million rate quotes per day, one might find something useful. In order to make the business operations more efficient, we analyse and compare price quote data and reservations history to find answers to several significant questions.

The company is interested in the following. How the number of reservations and quotes differ by year/season? How long in advance are prices checked and also how long in advance are reservations made? To see if there are unexpected changes (peaks, drops) in the amount of reservations and price quotes. See how the ratio of rate quotes to reservations differ by reseller, car class, seasons, customer's country and locations. To find out what is the actual demand after removing duplicate requests. One of the larger goals is also to find out if there is a way to use rate quotes to reservation ratio in order to derive historical demand based on reservations history.

The success of this project can be measured by how well it gives answers and how insightful they are to company's interests, which are described above.

Inventory of resources

Several laptops, hard drives, R, Python, global web resources, two master's students with at least 50 hours of time. Also there is a possibility to use University of Tartu's cluster to do some tasks which require more memory and computing capacity.

External knowledge by contacting CEO of RateChain, who has knowledge of the business. If needed then knowledge from data mining course's assistants and lecturer.

Requirements, assumptions, and constraints

The project must be completed by the poster session of Data Mining's course, which is on 8th of January. The data given to the students must not be leaked or given to third parties without permission.

The completed project must include a repository of code, which could be used to reproduce the results of the project, which gives answer to at least few to the questions described in the business goals section. Also a poster must be made which has to be presented during the poster session.

Risks and contingencies

We don't think there are any reasonable risks by working in a student team with two people. The only causes for delay could be that someone gets really sick or is not willing to participate. For the first one the other will try to cover for the other as well, but hopefully it does not happen. In the second case it probably will not happen, as we work very well together.

Terminology

Price Quote – Given input parameters such as age, location, source country and time period to the system, price quote is the cost the customer has to pay.

Reservation – When a price quote has been accepted by the customer then a reservation will be made. A car reserved for a customer for a certain time period and which can either be cancelled, confirmed or completed.

Duplicated request – ? (Not yet clarified)

Model – "A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions." (Oxford dictionary)

Costs and benefits

Not applicable.

Data-mining goals

By processing the given “rate_quotes” and “reservations” data set we give insights to the data according to the described business goals. The end goal is to make a poster about the results, which we have achieved and also present it in the poster session. The goal is to also have results together with code, which client can use to reproduce the results or use in the future for further analysis.

Data-mining success criteria

Whether the client is satisfied with the answers, such as what portion of price requests could be cached according to certain criterias. Graphs to show changes in the number of price quotes and reservations between different seasons, locations, source countries et cetera. Also answers to other interesting questions, such as how long in advance do customers check for prices and how long in advance reservations are made.

For question: “Can we use rate quotes to reservation ratio to derive historical demand based on reservations history?”, there will be an attempt to make a model to give approximations to the given question with no specific accuracy stated.

EX2. Data understanding

Outline data requirements:

Requirements for the data for our data mining goals are rather simple. For each rate quote and reservation we need to know the following:

- the timestamp when the rate quote or reservation was made;
- the time period for when the vehicle is rented out;
- the pickup and return location of the vehicle;
- the source country;
- the age of the customer;
- the class of the vehicle;
- the reseller of the vehicle.

Timestamp data should contain date (day, month, year) and time (at least hourly). For timestamps at least 2 year data is required in order to analyse whether reservations (bookings) are based on seasons or years.

Verify data availability

We confirm that most of the required data exists and is available to us. There are some issues with the data though. In the rate quote data instead of source country names, we have source country region ids, but for reservations we have source country names. Also the class of the vehicle is undefined for the rate quotes, instead we have a variable called car_class_example_id and for the reservations we have nothing about the car class. In order to make sense of this we will contact the client.

Define selection criteria

We have two csv-files (tables) Rate_quote.csv (23.44 GB) and Reservations_view (31.9 MB) containing . The data size is relatively high and therefore we might run into issues of lacking memory. To reduce this problem we will firstly develop and perfect everything on a subset of a data and later apply it on the whole data.

Describing data

1) Rate_quote.csv

Relevant fields:

- Uuid - unique ID for rate quote (character)
- Timestamp - timestamp for the rate quote (date)
- Pickup_desk_id - pickup location (integer)
- Return_desk_id - return location (integer)
- Car_class_example_id - don't know whether we will use it yet
- Pickup_timestamp - pickup date and time (date)
- Return_timestamp - return date and time (date)
- Broker_contract_id - reseller contract id (integer)
- Source_country_region_id - source country id (integer)
- Driver_age - driver age in rate request (integer)
- Vehicle_id - refers to row on price list used for price calculation (integer)

2) Reservations_view.csv

Relevant fields:

- Rate_quote_uuid - Unique ID for rate quote (character)
- Reservation_request_timestamp - reservation timestamp (date)
- Pickup_desk_id - pickup location (integer)
- Return_desk_id - return location (integer)
- Pickup_timestamp - pickup date and time (date)
- Return_timestamp - return date and time (date)
- Broker_reference - reseller id (integer)
- Source_country - customer country of residence (character)
- Driver_age - driver age (integer)
- Vehicle_id - refers to row on price list used for price calculation (integer)

As mentioned under "verify data availability" some of the variables don't have the same values. For example we do not have the ids for source countries. In rate quote data table we do have the source country region ids but not the source country name as in reservation data table.

Exploring data

We take a subset of 1 million rows of rate quote data and look at the range of the values and distributions. The variable **uuid** for rate quote is a character that consists of 36 characters (random letters and numbers). The variable **timestamp** is a date variable and the first million rate quotes (rows) have been made from 2017/12/01 to 2017/12/03. Variable **pickup_desk_id** and **return_desk_id** takes values from 1-5. The distribution of these variables are as following.

	1	2	3	4	5
pickup_desk_id	69.81 %	11.88 %	17.84 %	0.12 %	0.35 %
return_desk_id	69.35 %	12.38 %	17.68 %	0.11 %	0.47 %

Pickup_timestamp is a date variable and takes values from 2017/12/01 to 2018/12/26 and **return_timestamp** takes values from 2017/12/02 to 2019/01/02. **Broker_contract_id** and **source_country_region_id** are simply integers (id variables). **Driver_age** is a integer variable describing the age. The youngest driver is 19 and oldest 77, the mean is 30 years. **Vehicle_id** is an integer that takes values from 1 to 27. Similar variables are in the reservation data table and I will not bring out the distributions of the same variables. However let's look at the variable **source_country**. Three most frequently represented source countries (in random order) are UK, USA and South Korea.

Verifying data quality

We have found that most of the data we need indeed exists. Some of the data needs cleaning and reformatting as for example the dates are in different formats so we need to make sure that they are converted to one format. Also there are still some questions which we have about the data, which can be clarified by asking the client: what do some values mean in some columns.

There is also a lot of data we do not need in order to answer the questions posed. In general we have found that the data is really neat but needs a little bit extra information from the client.

EX3. Setting up and planning your project

Github: https://github.com/Abercus/dmproj2017

Project Plan

- Cleaning up the data
- Making sense of questions posed by the client
- Contacting the client
- Further exploration of the data
- Solving subtasks
    - To which time period customers are looking for rental cars now/yesterday/last week/last month?
    - How many days in advance customers are checking prices? How it changes through seasons and years?

- How many days/weeks/months in advance bookings are made? Does it vary based on season and years?
- Are there unexpected changes (peaks, drops) in bookings or price requests?
- How many price requests could be cached (have same pickup and return location, source country and driver age) for 1 hour, 3 hours, 12 hours?
- What is actual demand after eliminating duplicated requests?
- What is "rate quotes to reservation" ratio by resellers? Does it vary by seasons or source countries?
- What is "rate quotes to reservation" ratio by car class and location?
- Can we use rate quotes to reservation ratio to derive historical demand based on reservations history?

- Graphical illustrations
- Creating a poster for the poster session

Time estimations

Nr	Task to be done	Estimated time
1	Cleaning data	2h - Pirge
2	Clarifying questions posed by client	2h - Joonas
3	Further exploration of the data	4h – 2h Pirge, 2h Joonas
4	Solving subtasks	30h – 15h Pirge, 15h Joonas
5	Graphical illustrations	7h – 4h Pirge, 3h Joonas
6	Creating a poster	5h – 2h Pirge, 3h Joonas