A PROJECT IN DATA SCIENCE

The return at the Yeti

The problem

The new client assigned to Angela, an analyst at the renowned consulting firm GNRL, corresponds to the travel agency Yeti-Travel. The agency has detected a noticeable drop in sales during the last year, especially among its regular customers. The situation represents a serious problem for the agency, considering that a large part of the sales comes from regular customers. Moreover, over the years, the agency has become specialized in one specific type of service: the sale of travel packages for schools. For this reason, the loss of its regular customers threatens the continuity of the company.

After a brief conversation with the client, Angela believed that the best way to tackle the problem was to build a predictive model that uses historical information of loyal customers during a particular year to identify their more relevant features.

Angela decides to organize her work in three different stages: (1) the gathering of internal/external data, (2) the development of a model to predict the churn probability, and finally (3) the analysis of the most informative features for the design of specific measures to strengthen customer loyalty. For example, to establish a deeper relationship with some of the clients or to design customized products for them based on their previous preferences.

The company

The Yeti Travel Agency has a long tradition, born in the mid 2000s the company quickly became an important player in the market. Despite its rapid expansion, the company has retained its family character with all the advantages and disadvantages that this entails. In the beginning, the company offered multiple services but quickly specialized in a specific part of the market: school trips.

One of the main characteristics of the company is its wide range of trips with alternatives that include trips of different scales and durations. Moreover, since final users are very young, the planning process is treated with special attention and anticipation by the company. The agency has designed a well-defined structure that includes the participation of some teachers and/or guardians in the trip, for which the agency offers a discounted price. Additionally, the agency organizes several meetings with the schools to fine-tune details and solve potential problems.

The Data

The company is aware of the importance of historical information and keeps track of a big part of the process, nevertheless, the information is not centralized but split between the different sales agents and the different areas of the company.

Angela decided to collect all the available data and make a snapshot of clients in two consecutive years. In this manner, she can create a model to predict if a client came back based on the information in the snapshot taken at the end of the first year. Before the development of a predictive model, she should gather information from different departments that, on many occasions, was not standardized, incomplete, and/or inconsistent.

After an evaluation of the availability and the potential predictive power of the information, angela together with the client and the strategy team decided to use the data of three different divisions: (a) Sales, (b) Finance, and (c) CRM. The dataset information is detailed in the tables annexed to this document.

You are provided with two zip files, one containing past, labeled data and one containing new unlabeled data.

Submission: requirements and procedure

Requirements

• Vector of predictions for the unlabeled dataset
  - .csv file
  - One row per obs, as many rows as the unlabeled set
• Slides of your presentation
  - .pdf file
• Source code to reproduce the results you obtained. This is the code that you used to obtain your model.

Procedure

The submission procedure consists of an email for each group to andrea.mor@polimi.it. The email must contain **three** attachments:

• surname1_surname2_surname3.csv (the predictions)
• surname1_surname2_surname3.pdf (the slides)
• surname1_surname2_surname3.[ipynb/py/etc.] (the code)

The **deadline** for the submission is **17/04/2023, 16:00 CET**.

Sales Dataset

Data field	Description
ID	ID sales department.
Program_Code	Program code of the trip.
From_Grade	Lowest grade in school of a participant.
To_Grade	Highest grade in school of a participant.
Group_State	School location.
Days	Number of days on the program.
Travel_Type	Travel mode (A = Air, B = Bus, T = Train).
Departure_Date	Departure day.
Return_Date	Return day.
Early_RPL	First communication date inviting people to join.
Latest_RPL	Last communication inviting people to join.
Cancelled_Pax	Number of passengers who made a deposit but cancelled.
Total_Discount_Pax	Number of extra passengers (e.g. professors)
Initial_System_Date	First date when trip was organized.
SPR_Product_Type	Aggregation of tour types.
FPP	Number of full-payment participant.
Total_Pax	Number of total passengers (including extra participants).
DepartureMonth	Month of departure.
GroupGradeTypeLow	Lowest grade type in the trip.
GroupGradeTypeHigh	Highest grade type in the trip.
GroupGradeType	Combination of the above.
MajorProgramCode	Aggregation of program codes.
FPP_to_School_enrollment	The ratio of FPP to school enrollment.
Retained	target

Finance Dataset

Data field	Description
ID	ID finance department.
Deposit_Date	Expected deposit day.
Special_Pay	Payment modality (internal code)
Tuition	Price per full-payment participant (FPP).
FRP_Active	Number of FPPs who bought trip-cancellation insurance.
FRP_Cancelled	Number of FPPs who bought trip-cancellation insurance and cancelled it.
FRP_Take_up_percent_	Percentage of FPPs who bought the insurance pay for it.
EZ_Pay_Take_Up_Rate	Percentage of FPPs use automatic bank draft.
School_Sponsor	Indication of whether or not the school is sponsoring the trip.
SPR_Group_Revenue	Amount paid for all of the participants.
FPP_to_PAX	Percentage of FPP.
Num_of_Non_FPP_PAX	Number of non-FPP participants.

CRM Dataset

Data field	Description
ID	ID CRM department.
Poverty_Code	Poverty code for the school area based on estimated percentage below the poverty line. A is 0 to 5.9, B is 6 to 15.9, C is 16 to 30.9, D is 31 or more, E is unclassified, Space if DISTCLASS = U (Supervisory Union).
Region	State areas.
CRM_Segment	CRM code system (internal code)
School_Type	Public or private.
Parent_Meeting_Flag	Indication whether a parent meeting was held.
MDR_Low_Grade	Lowest grade (not just participants) in the school.
MDR_High_Grade	Highest grade (not just participants) in the school.
Total_School_Enrollment	School enrollments.
Income_Level	Parent income level code. A is lowest, Q is highest, Z is unclassified.
SPR_New_Existing	New client indicator.
NumberOfMeetingswithParents	Number of meetings with parents prior to the trip.
FirstMeeting	The date of the first meeting with parents (NA if none held).
LastMeeting	Date of the last meeting with parents (NA if none held).
DifferenceTraveltoFirstMeeting	Days from the first parent meeting to travel date.
DifferenceTraveltoLastMeeting	Days from the last parent meeting to travel date.
SchoolGradeTypeLow	The lowest grade type in the school.
SchoolGradeTypeHigh	The highest grade type in the school.
SchoolGradeType	Combination of the above denoting the type of school.
SchoolSizeIndicator	Size of the school (S, M, L, S-M, M-L).