2nd Datathletes (Q1- 2018)

As a part of the ongoing Data Analytics Challenge at Amadeus Labs, we are happy to bring to you the 2nd edition of the Datatheletes Challenge (Q1-2018).

Fare Choice Prediction Challenge

Overview

Can you predict the fare choice for a customer using their booking data?

When it comes to booking a flight, not all customers have the same needs. Some customers desire a complimentary meal, others want to travel light and on a budget, whilst others want to be able to make last minute travel decisions or have free cancellation, carry extra baggage, earn more air miles, enjoy benefits like lounge access, upgrade eligibility and priority check-in. Airlines try to design fare plans which include some of these privileges to provide customers more choice, customisation and transparency for their travel requirements.

Data

Description

The data for this contest has all the economy class one way, non-stop flights for some of the top domestic routes in India for June 2017 for an airl economy are Classic, Deal, Flex and Saver.

Data fields

The following information is included:

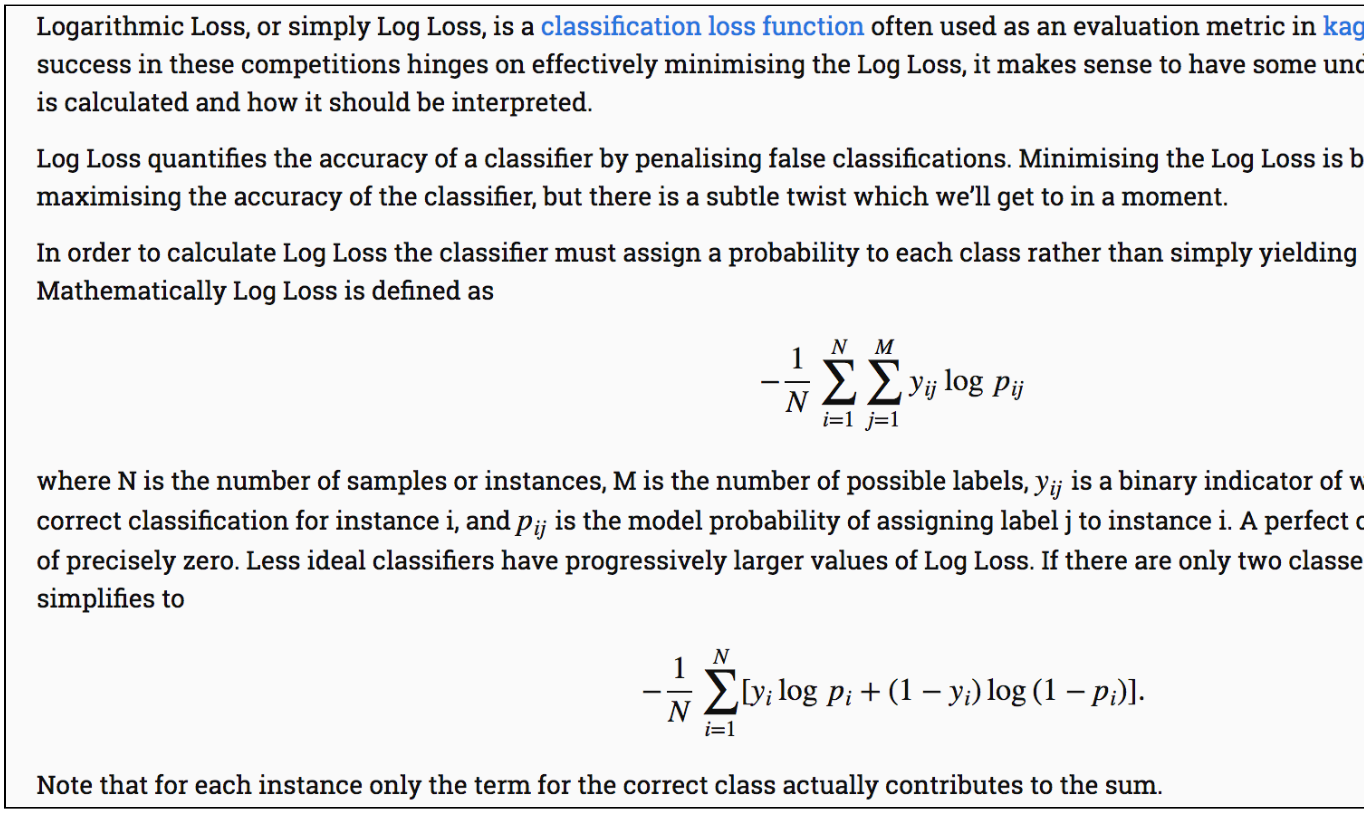
- id - row number
- booking_date – date of booking the ticket (YYYYMMDD)
- origin – origin airport code
- destination – destination airport code
- dep_date – departure date of the flight (YYYYMMDD)
- dep_time – departure time of the flight
- pax – number of passengers in the PNR ticket
- fare_choice – only in train.csv, this is the class to be predicted

File descriptions

- train.csv - the training set, contains booking data with their fare choices
- test.csv - the test set, you must predict probability of all fare choices for each of the bookings
- sample_submission.csv - a sample submission file in the correct format

Evaluation

- Submissions are evaluated on multi-class logarithmic loss between the predicted probability and the observed target. **Lower values for log-** Brief explanation on **multi-class logarithmic loss:**



Logarithmic Loss, or simply Log Loss, is a classification loss function often used as an evaluation metric in kag
success in these competitions hinges on effectively minimising the Log Loss, it makes sense to have some und
is calculated and how it should be interpreted.

Log Loss quantifies the accuracy of a classifier by penalising false classifications. Minimising the Log Loss is b
maximising the accuracy of the classifier, but there is a subtle twist which we'll get to in a moment.

In order to calculate Log Loss the classifier must assign a probability to each class rather than simply yielding
Mathematically Log Loss is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\, p_{ij}$$

where N is the number of samples or instances, M is the number of possible labels, y_{ij} is a binary indicator of w
correct classification for instance i, and p_{ij} is the model probability of assigning label j to instance i. A perfect c
of precisely zero. Less ideal classifiers have progressively larger values of Log Loss. If there are only two classe
simplifies to

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log\, p_i + (1-y_i)\log\,(1-p_i)].$$

Note that for each instance only the term for the correct class actually contributes to the sum.

Submission File

- For each booking in the test set, you must predict a probability for each of the different classes. The file should contain a header and have th sample-submission file. The order of classes in the file should not be changed. The class names are case-sensitive.

Competition Rules

1. The competition will be in two stages. In the first stage, the position in the leaderboard will be used to select the top teams. The number of su restricted to 15. The best submission for each team will be used to determine the rankings. The leader board will be computed on 100% of th
2. In the final round, the selected teams will be asked to make a brief presentation (not more than 3 slides) on their approach and any additiona performance in the leaderboard as well as the presentation will be evaluated to decide the final winners.
3. The selected teams must be able to reproduce their best result on leaderboard during their demo presentation; hence they should keep a tra submission scores.

> Sample Submission File and Train / Test sets can be found attached here.