Revenue

Hide

Task

Medaplexis leases surgical instruments to medical facilities and has contracts with hospitals in several districts. The company wants to utilize predictive analytics to better forecast demand for manufacturing and inventory. They're also interested in forecasting the annual revenue that will be generated by different districts.

Medaplexis wants to use their data related to *hospital profile*, *current revenue*, and *projected revenue* to predict the likelihood of each hospital leasing each instrument and generating revenue at a district level.

Given Medaplexis' dataset, predict whether or not a hospital will lease an instrument and, if it does lease the instrument, how much revenue will it generate?

Dataset

Download the zip file using this link. The dataset contains files named *HospitalProfiling.csv*, *HospitalRevenue.csv*, *ProjectedRevenue.csv*, and *Solution.csv*. The files are organized as follows:

- **HospitalProfiling.csv**: This dataset contains hospital profiles. The file is organized into the following fields:
  - *Hospital_ID* is the hospital's unique ID.
  - *District_ID* is the ID of district where the hospital is located.
  - *Hospital_employees* is the total number of permanent employees at the hospital.
- **HospitalRevenue.csv** - This dataset contains hospital revenue for the current year. The file is organized into the following fields:
  - *Hospital_ID* is the hospital's unique ID.
  - *Region_ID* is the region ID for the region where the hospital is located.
  - *District_ID* is the ID of district where the hospital is located.
  - *Instrument_ID* is the ID of the instrument that the hospital is leasing.
  - *Transaction_Revenue* is the revenue (in *dollars*) generated at the end of the month from related tractions. The transaction revenue is captured for 12 months and provided in 12 separate fields named *Month 1*, *Month 2*, *Month 3*, ..., *Month 12*. There is also an additional field named *Year Total* which contains the sum of the revenues over 12 months.
- **ProjectedRevenue.csv** - This dataset provides details about leasing deals closed by hospitals which should show up as hospital revenue for the next year. The file is organized into the following fields:
  - *Hospital_ID* is the hospital's unique ID.
  - *District_ID* is the ID of district where the hospital is located.
  - *Instrument_ID* is the ID of the instrument that the hospital *is* leasing.
  - *Annual_Projected_Revenue* is the projected revenue (in *dollars*) for the next year.

The **Solution.csv** is provided as an output file. The file is organized into the following fields:

- *Hospital_ID* is the hospital's unique ID.

- *District_ID* is the ID of district where the hospital is located.
- *Instrument_ID* is the ID of the instrument that the hospital *might* lease.
- *Buy_or_not*
- *Revenue*

The fields *Buy_or_not* and *Revenue* are empty. You should predict the values for these fields for each combination of *Hospital_ID*, *District_ID*, and *Instrument_ID* without editing the existing values in the file. Note that:

- *Buy_or_not* is a binary flag where 0 denotes that a hospital in a district will *not* lease a given instrument ID and 1 denotes that a hospital in a district *will* lease a given instrument ID.
- *Revenue* is the revenue (in *dollars*) generated from each combination of *Hospital_ID*, *District_ID*, and *Instrument_ID*.

Submission Details
Upload the following three files:

- The output file *Solution.csv* (max allowed size is *10MB*) containing the predicted values. Make sure that the order of the fields in the output file is the same (i.e., *Hospital_ID*, *District_ID*, *Instrument_ID*, *Buy_or_not*, and *Revenue*). *Do not* remove the field headers from the output file.

  A valid *Solution.csv* has the following format:

  ```
  Hospital_ID,District_ID,Instrument_ID,Buy_or_not,Revenue
  Hospital 1,District 39,Instrument 3,1,89866
  Hospital 1,District 50,Instrument 15,0,0
  Hospital 1000,District 16,Instrument 11,1,59718
  Hospital 1000,District 16,Instrument 15,1,43760
  Hospital 1000,District 18,Instrument 11,0,0
  Hospital 1000,District 18,Instrument 15,1,35516
  Hospital 1000,District 28,Instrument 11,0,0
  Hospital 1000,District 28,Instrument 15,1,68978
  Hospital 1000,District 37,Instrument 11,0,0
  Hospital 1000,District 39,Instrument 11,1,88073
  ```

- A *PDF* file (max allowed size is *1MB*) providing the findings and justification on the following topics:
  - Write a few lines about training dataset quality and any errors found in the training dataset.
  - Explain the data preprocessing steps.
  - Justify the model chosen by you for the prediction.
- The source code is written for the training and prediction. Upload a *zip* file (max allowed size is *2MB*). The submitted file must have a *README* file with a detailed description about how to run the model to predict the missing values and generate the *Solution.csv* file.

  There is no limit on execution time, but the code should generate the output file: *Solution.csv*. Only *open source languages* are allowed.

Evaluation
The following evaluation schemes will be used for evaluating prediction accuracy:

- *Buy_or_not* Prediction Accuracy:
  Let T_P be the *true positives*, F_P be the *false positives*, T_N be the *true negatives* and F_N be the *false negatives*. Now we define precision P and recall R:

$$P = \frac{T_P}{T_P + F_P}$$

$$R = \frac{T_P}{T_P + F_N}$$

Now we will calculate the accuracy for *Buy_or_not* prediction:

$$S_{Buy\_or\_not} = \frac{2PR}{P + R}$$

- *Revenue* Prediction Accuracy:
  Let $R_{expected}$ be the value of the annual revenue in the evaluation dataset and $R_{predicted}$ be the predicted value of the annual revenue. If the total number of predicted values is N, then we can calculate the *model mean square error (MMSE)* and *base mean square error (BMSE)*:

$$MMSE = \frac{1}{N}\left(\sum_{i=1}^{N}\left(R_{predicted_i} - R_{expected_i}\right)^2\right)$$

$$BMSE = \frac{1}{N}\left(\sum_{i=1}^{N}\left(R_{expected_i} - \mu\right)^2\right)$$

  Here, $\mu = \frac{1}{N}\left(\sum_{i=1}^{N} R_{expected_i}\right)$. Now we will calculate the accuracy for *Revenue* prediction:

$$S_{Revenue} = \begin{cases} 0 & MMSE > BMSE \\ \frac{BMSE - MMSE}{BMSE} & Otherwise \end{cases}$$

Both the prediction accuracy scores, $S_{Buy\_or\_not}$ and $S_{Revenue}$, are normalized to $[0, 1]$. The accuracy of the prediction model is given by S:

$$S = 0.5 \times S_{Buy\_or\_not} + 0.5 \times S_{Revenue}$$

Ranking

During the contest, we will evaluate the accuracy of the prediction model on 50% of pre-sampled dataset. After the contest, we'll perform a full evaluation on the remaining 50% of the dataset. This final evaluation will only be performed on your *last* uploaded output file (i.e., most recently uploaded), so be sure that your final submission is the best output file (i.e., the file having the largest score S).

File Upload