Supervised Learning Competition

Rules:

- Teams of maximum 3 people [recommended]. Doesn't have to be your project team.
- You can discuss general ideas with other teams but not repeat their exact analytics process.
- One submission per team.
- Don't use any external data sources.
- Your submission MUST include:
    i. A 1-2 pages report stating: (1) Team members' names and student IDs, (2) the used software packages, (3) instructions on how to download and install them, (4) a detailed description of the analytics process (Exploration, preparation and modeling) used with justification for each step, (5) a brief description of the other approaches you tried but didn't work out and (6) the accuracy you achieved (Must show the confusion matrix on validation data).
    ii. All code developed to produce the predictions.
    iii. The unlabeled file with an added column representing the predicted value (i.e the results from running your analytics process on the unlabeled file).
- Create a folder for each task and compress all folders into one archive that you upload to OnQ.
- Late submissions will be penalized 1 point for each late day.

Only solve one of these two tasks.

Task 1: Predict patients readmission to the hospital

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks[1]. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

- It is an inpatient encounter (a hospital admission).
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

Attribute Information:

(1) **Encounter ID** *Numeric* Unique identifier of an encounter

(2) **Patient number** *Numeric* Unique identifier of a patient

(3) **Race** *Nominal* Values: Caucasian, Asian, African American, Hispanic, and other

(4) **Gender** *Nominal* Values: male, female, and unknown/invalid

(5) **Age** *Nominal* Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)

(6) **Weight** *Numeric* Weight in pounds.

(7) **Admission type** *Nominal* Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available

(8) **Discharge disposition** *Nominal* Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available

(9) **Admission source** *Nominal* Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital

(10) **Time in hospital** *Numeric* Integer number of days between admission and discharge

(11) **Payer code** *Nominal* Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay

(12) **Medical specialty** *Nominal* Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

(13) **Number of lab procedures** *Numeric* Number of lab tests performed during the encounter
(14) **Number of procedures** *Numeric* Number of procedures (other than lab tests) performed during the encounter
(15) **Number of medications** *Numeric* Number of distinct generic names administered during the encounter
(16) **Number of outpatient visits** *Numeric* Number of outpatient visits of the patient in the year preceding the encounter
(17) **Number of emergency visits** *Numeric* Number of emergency visits of the patient in the year preceding the encounter
(18) **Number of inpatient visits** *Numeric* Number of inpatient visits of the patient in the year preceding the encounter
(19) **Diagnosis 1** *Nominal* The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
(20) **Diagnosis 2** *Nominal* Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
(21) **Diagnosis 3** *Nominal* Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
(22) **Number of diagnoses** *Numeric* Number of diagnoses entered to the system
(23) **Glucose serum test result** *Nominal* Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured
(24) **A1c test result** *Nominal* Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
(25) **Change of medications** *Nominal* Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"
(26) **Diabetes medications** *Nominal* Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
(27 – 49) **23 features for medications** *Nominal* For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed
(50) **Readmitted** *Nominal* Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.

<u>Objective:</u>

- Use the *C2T1_Train.csv* file to train a predictive model that can predict if a patient will be readmitted to the hospital again or not.
- **Bonus (1 point):** Train your model to predict if a patient will be readmitted less than or more than 30 **<u>DAYS</u>**.
- Once you have your model trained, predict the readmission of the patients in the *C2T1_Test.csv* file.
- Create a labeled *C2T1_Test_Labled.csv* file using your predictive model to predict the readmission values for the patient records in the C2T1_Test.csv file. Your *C2T1_Test_Lableled.csv* file should look like this:

```
encounter_id,patient_nbr, readmitted
168899772,88565423, No
168903044,61086362, <30
168903046,61086362, >30
```

- Accuracy is measured as (number of correct predictions)/(Total number of records).

Task 2: Predict number of comments a Facebook post receives

The leading trends towards social networking services had drawn massive public attention. It created new ways for sharing news, advertisements and marketing. In this task, you are given a dataset[2] of some posts that appeared in the past, whose target values (comments received) are already known. The task is to predict how many comments that a post is expected to receive in the next H hrs. The data originates from Facebook pages. The raw data is crawled by crawler and cleaned on basis of following criteria:

- Only those posts that were published in last three days w.r.t to the base date/time are considered as it is expected that the older posts usually don't receive any more attention.
- Posts with missing details were omitted.

Attribute Information:

(1) **PageLikes** *Numeric* Defines the popularity or support for the source of the document.

(2) **PageCheckIns** *Numeric* Describes how many individuals so far visited this place. This feature is only associated with the places eg:some institution, place, theater etc.

(3) **DailyInterest** *Numeric* Defines the daily interest of individuals towards source of the document/ Post. The people who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares, etc by visitors to the page.

(4) **PageCategory** *Nominal* Defines the category of the source of the document eg: place, institution, brand etc.

(5-9) **PageCC1Min, PageCC1Max, PageCC1Avg, PageCC1Median, PageCC1Std** *Numeric* These features are aggregated by page, by calculating min, max, average, median and standard deviation of attribute: TotalComments_CC1.

(10-14) **PageCC2Min, PageCC2Max, PageCC2Avg, PageCC2Median, PageCC2Std** *Numeric* These features are aggregated by page, by calculating min, max, average, median and standard deviation of attribute: CommentsLast24H_CC2.

(15-19) **PageCC3Min, PageCC3Max, PageCC3Avg, PageCC3Median, PageCC3Std** *Numeric* These features are aggregated by page, by calculating min, max, average, median and standard deviation of attribute: CommentsLast48to24H_CC3.

(20-24) **PageCC4Min, PageCC4Max, PageCC4Avg, PageCC4Median, PageCC4Std** *Numeric* These features are aggregated by page, by calculating min, max, average, median and standard deviation of attribute: CommentsFirst24H_CC4.

(25-29) **PageCC5Min, PageCC5Max, PageCC5Avg, PageCC5Median, PageCC5Std** *Numeric* These features are aggregated by page, by calculating min, max, average, median and standard deviation of attribute: CC2MinusCC3_CC5.

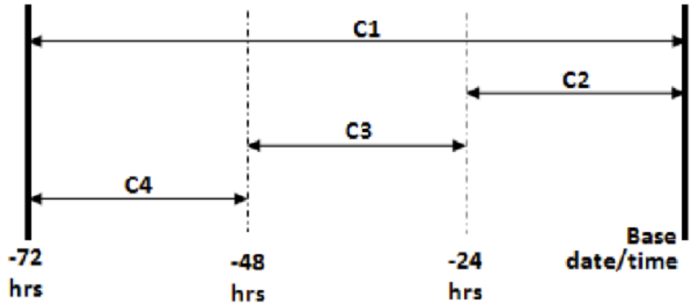
(30) **TotalComments_CC1** *Numeric* Total comment count before selected base date/time.

(31) **CommentsLast24H_CC2** *Numeric* Comment count in last 24 hrs w.r.t toselected base date/time.

(32) **CommentsLast48to24H_CC3** *Numeric* Comment count is last 48hrs to last 24 hrs w.r.t to base date/time.

(33) **CommentsFirst24H_CC4** *Numeric* Comment count in first 24 hrs after publishing the post.

(34) **CC2MinusCC3_CC5** *Numeric* The difference between CC2 and CC3.



[2] Kamaljot Singh, Ranjeet Kaur, and Dinesh Kumar. 2015. Comment Volume Prediction Using Neural Networks and Decision Trees. In Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation (UKSIM '15). IEEE Computer Society, Washington, DC, USA, 15-20. DOI=http://dx.doi.org/10.1109/UKSim.2015.20

(35) **TimeSincePublishedinHrs** *Numeric* Number of hours since post was published. Range (0-71)

(36) **PostLength** *Numeric* Character count in the post.

(37) **PostShareCount** *Numeric* Number of post shares, that's how many peoples had shared this post on to their timeline.

(38) **PostPromoted** *Binary* whether the post is promoted (1) or not (0).

(39) **PredictAfterHrs** *Numeric* Number of hours after which you should predict the number of comments.

(40-46) **Published_Sunday, Published_Monday, Published_Tuesday, Published_Wednesday, Published_Thursday, Published_Friday, Published_Saturday** *Binary* This represents the day (Sunday...Saturday) on which the post was published.

(47-53) **PredictOn_Sunday, PredictOn_Monday, PredictOn_Tuesday, PredictOn_Wednesday, PredictOn_Thursday, PredictOn_Friday, PredictOn_Saturday** *Binary* This represents the day (Sunday...Saturday) on which the you are to predict the number of comments.

(54) **CommentsNumber** *Numeric (Target)* Number of comments the post received after the *PredictAfterHrs* hours.

Objective:

- Use the *C2T2_Train.csv* file to train a predictive model that can the number of comments a post will receive after *PredictAfterHrs* hours.
- **Bonus (1 point):** Convert this problem to a classification problem and train another model.
- Once you have your model trained, predict the number of comments for the posts in the *C2T2_Test.csv* file.
- Create a labeled *C2T2_Test_Labled.csv* file using your predictive model to predict the number of comments for the posts in the C2T2_Test.csv file. Your *C2T2_Test_Lableled.csv* file should look like this:

```
ID, CommentsNumber
1, 16
2, 0
3, 150
```

- Accuracy is measured using Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$