Work Integrated Learning Programmes | BITS Pilani Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science and Machine Learning Workshop at Intuit

(2nd sem, 2019-20)

1. Objective

The workshop aims to provide students with a platform for guided problem solving in data science using feature engineering techniques and machine learning, by applying the concepts learnt during the coursework on Introduction to Data Science and Machine Learning. This is also a graded evaluation component towards EC1 for these 2 courses.

2. Date, Time, Venue

- Feb 23, 2020 [Sunday], 830AM-530PM
- Location: Intuit

3. Topics Covered

- Feature engineering techniques
- Machine learning classifiers

4. Workshop Plan

a) Before the workshop

- You are divided into groups of 3-5.
- Each group would be given a problem statement, and would be expected to come in to the workshop with baseline performance figures for the dataset assigned to them.
- There will be 2 checkpoints for discussions/clarifications before the workshop.
- Submission: Brief summary of analysis and observations during the prep work [Template will provided]
- [Duration: 3 weeks]

b) During the workshop

- Session 1: Impact of feature engineering on ML classifiers [Duration: 2.5 hrs]
- Session 2: Application of ensemble techniques for classification [Duration: 2.5 hrs]
- Session 3: Student group presentations [Duration: 2 hrs]
- Summary and conclusion [Duration: 1 hr]

c) After the workshop

- Report submission based on activities carried out during the workshop and insights on the results obtained [Template will be provided]
- Duration: [1 week]

5. Evaluation Rubrics

Course	Prep Work	Results & presentation during the workshop*	Report
Introduction to Data Science	5	10	5
Machine Learning	5	15	5

*Attendance is mandatory for the workshop. There will be no makeup component for this. Individual marks may vary based on participation and contribution.

6. Important Dates

Milestone	Date
Allotment of problem statements, with datasets	Jan 16, 2020
Checkpoint 1	Feb 1, 2020
Checkpoint 2	Feb 15, 2020
Submission of prep work summary	Feb 21, 2020
Workshop	Feb 23, 2020
Report submission	March 1, 2020

7. Assignment of Problem Statements to Groups

See below for details on problem statements. The assignment of problem statements to groups is as follows:

Problem Statement 1: Group 7
Problem Statement 2: Group 8, Group 9
Problem Statement 3: Group 1, Group 5
Problem Statement 4: Group 2, Group 3
Problem Statement 5: Group 4, Group 6

8. Problem Statements

a) Detection of anomalies in credit card transactions

Objective:

- Detect anomalies in credit card transactions

Dataset:

- https://drive.google.com/drive/folders/1XCKR-Gdw9yziWWW6TVvx5wvArFkP96PP

Prep work required:

- Identification of the performance parameters to be improved
- Baseline performance figures for 5 different ML classifiers, after minimal data pre-processing
- Observations from exploratory analysis of the dataset. One of the classifiers must be ANN.
- Outline of feature engineering techniques that may be used to improve the classifier performance

Session 1:

- Shortlist 3 best performing classifiers, from the prep work
- Apply relevant feature engineering techniques on the dataset provided
- Compare the performance of the 3 classifiers with the baseline performance figured obtained during prep-work
- Note down the features that figure high in feature ranking

Session 2:

- Apply ensemble techniques to observe their impact on the performance of the classifiers

Session 3:

- Group presentation

b) Predictive loan model for an applicant

Objective:

- Predict the risk of a loan being default based on the past loan data, for a given loan applicant

Dataset:

- https://drive.google.com/drive/folders/1XCKR-Gdw9yziWWW6TVvx5wvArFkP96PP
- Dataset will be provided in parts

Prep work required:

Data for 3 quarters will be provided to the students, with a select subset of features. This will be used for prep work and initial baselining.

- Identification of the performance parameters to be improved, for the given problem statement
- Baseline performance figures for 5 different ML classifiers, after minimal data pre-processing, Baseline figures must include (i) accuracy (ii) classification report (iii) confusion matrix (iv) ROC-AUC and AUPRC scores. One of the classifiers must be ANN.
- Observations from exploratory analysis of the dataset
- Outline of feature engineering techniques that may be used to improve the classifier performance

Session 1:

The complete dataset (with data for 2 more years and all features) will be provided.

- Shortlist 3 best performing classifiers, from the prep work
- Apply relevant feature engineering techniques on the dataset provided
- Compare the performance of the 3 classifiers with the baseline performance figured obtained during prep-work
- Note down the features that figure high in feature ranking

Session 2:

- Apply ensemble techniques to observe their impact on the performance of the classifiers

Session 3:

- Group presentation

c) Fall detection

Objective:

- Detect falls based on the dataset comprising features from kinematic sensor parameters.

Dataset:

- https://drive.google.com/drive/folders/1XCKR-Gdw9yziWWW6TVvx5wvArFkP96PP

Prep work required:

- Identification of the performance parameters to be improved, for the given problem statement
- Baseline performance figures for 5 different ML classifiers, after minimal data pre-processing, Baseline figures must include (i) accuracy (ii) classification report (iii) confusion matrix an (iv) ROC-AUC and AUPRC scores. One of the classifiers must be ANN.
- Observations from exploratory analysis of the dataset
- Outline of feature engineering techniques that may be used to improve the classifier performance

Session 1:

- Shortlist 3 best performing classifiers, from the prep work
- Apply relevant feature engineering techniques on the dataset provided
- Compare the performance of the 3 classifiers with the baseline performance figured obtained during prep-work
- Note down the features that figure high in feature ranking

Session 2:

- Apply ensemble techniques to observe their impact on the performance of the classifiers

Session 3:

- Group presentation

d) Predicting music subscription cancellations

Objective:

- Predict if subscription users of a music streaming service will churn or stay after their current membership expires.

Dataset:

- https://drive.google.com/drive/folders/1XCKR-Gdw9yziWWW6TVvx5wvArFkP96PP

Prep work required:

- Identification of the performance parameters to be improved, for the given problem statement
- Baseline performance figures for 5 different ML classifiers, after minimal data pre-processing, Baseline figures must include (i) accuracy (ii) classification report (iii) confusion matrix an (iv) ROC-AUC and AUPRC scores. One of the classifiers must be ANN.
- Observations from exploratory analysis of the dataset
- Outline of feature engineering techniques that may be used to improve the classifier performance

Session 1:

- Shortlist 3 best performing classifiers, from the prep work
- Apply relevant feature engineering techniques on the dataset provided
- Compare the performance of the 3 classifiers with the baseline performance figured obtained during prep-work
- Note down the features that figure high in feature ranking

Session 2:

- Apply ensemble techniques to observe their impact on the performance of the classifiers

Session 3:

- Group presentation

e) Sentiment Analysis

Objective:

- Analysis of product reviews and customer ratings to classify whether the reviews and positive or negative.

Dataset:

- https://drive.google.com/drive/folders/1XCKR-Gdw9yziWWW6TVvx5wvArFkP96PP

Prep work required:

A subset of data will be given for initial analysis.

- Identification of the performance parameters to be improved, for the given problem statement
- Baseline performance figures for 5 different ML classifiers, after minimal data pre-processing, Baseline figures must include (i) accuracy (ii) classification report (iii) confusion matrix and (iv) ROC-AUC and AUPRC scores. One of the classifiers must be ANN.
- Outline of dictionaries that may be used to improve the classifier performance
- Word cloud for positive and negative sentiments

Session 1:

The complete dataset will be provided.

- Shortlist 3 best performing classifiers, from the prep work
- Use 3 relevant dictionaries based on words or bag of words and classify into 5 classes.
- Compare the performance of the 3 classifiers with the baseline performance figured obtained during prep-work
- Word cloud for positive and negative sentiments

Session 2:

- Build a dictionary form the given dataset and classify.
- Compare the performance of the 3 classifiers
- Word cloud for positive and negative sentiments

Session 3:

- Group presentation