CSE 351 - Introduction to Data Science (Summer 2020)
Prof. Praveen Tripathi
Course Projects

Soft Deadline for EDA part: June 23
(You don't need to make a submission for the soft deadline, but you should report to the mentor on that day through office hours or emails.)
Blackboard Submission Deadline: June 29 11:59 pm (one submission per team)
Zoom Presentations: June 30, July 2, July 3

Do your own work!

Choice 1: Fatal Force in the US

Mentor: Yunting Yin <yunyin@cs.stonybrook.edu>

In the United States, use of deadly force by police has been a high-profile and contentious issue. 1000 people are shot and killed by US cops each year. The ever-growing argument is that the US has a flawed Law Enforcement system that costs too many innocent civilians their lives. In this project, we will analyze one of America's hottest political topics, which encompasses issues ranging from institutional racism to the role of Law Enforcement personnel in society.

Datasets:
The "fatal_force.zip" file contains six datasets to use for this project. "police_killings_train.csv" and "police_killings_test.csv" are mandatory datasets with self-explanatory data fields. The other four files "share_race_by_city.csv", "income.csv", "poverty.csv", and "education.csv" are optional datasets you can use to perform analysis and to add features to your models.

The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:
EDA (10%):
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Clean and merge the datasets, explain what you did.
- Which state has the most fatal police shootings? Which city is the most dangerous?
- What is the most common way of being armed?
- What is the age distribution of the victims? Compare age distribution of different races.
- Compare the total number of people killed per race. Compare the number of people killed per race as a proportion of respective races. What difference do you observe?

(Note: EDA should be performed on the training set only)

Modeling and Question Answering (10%):
Apply three machine learning algorithms to explore whether it is possible to predict the race of a victim based on other features. Train your models on the training set, and make predictions for the test set with the “race” column dropped. Evaluate the accuracy of your predictions. If your predictions are not very accurate, what do you think is the reason?

Project Report (10%):
Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):
Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file/notebook as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.

Choice 2: What happens to animals in shelter?

Mentor: Yunting Yin <yunyin@cs.stonybrook.edu>

Every year, approximately 7.6 million companion animals end up in US shelters. Many animals are given up as unwanted by their owners, while others are picked up after getting lost or taken out of cruelty situations. Many of these animals find forever families to take them home, but just as many are not so lucky. 2.7 million dogs and cats are euthanized in the US every year. In this project, we analyze trends and make predictions for outcomes of shelter animals.

Datasets:
The “shelter_animals.zip” file contains the datasets for this project. All data fields are self-explanatory. “OutcomeType” is the label we are trying to predict.

The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:

EDA (10%):
Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Normalize the age feature, which is originally in different units.

- Some features have a large number of unique values, find a way to categorize them into groups.
- Explore how different outcomes are distributed for cats and dogs.
- What time in the day are the animals adopted/transferred/euthanized? Is there a pattern?
- What features are the most important ones in determining the outcome?

(Note: EDA should be performed on the training set only)

Modeling and Question Answering (10%):

Build three models, train them on the training set, and predict the outcome for the test set (after dropping the “OutcomeType” column in the test set). Explain how each model works (briefly introduce the machine learning algorithms behind them). Evaluate the performance of each model based on the original outcome in the test set. Which model works best?

Project Report (10%):

Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):

Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file/notebook as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.

Choice 3: What makes people in a country happy?

Mentor: Yunting Yin <yunyin@cs.stonybrook.edu>

The World Happiness Report is a landmark survey of the state of global happiness that ranks countries by how happy their citizens perceive themselves to be. The report gains global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. This project allows us to gain insight into the state of happiness in the world today.

Datasets:

The “world_happiness.zip” file on Blackboard contains happiness data for different countries from year 2015 to year 2019. We will treat data of year 2015 to year 2018 as the training set, and year 2019 data as the test set. Description of the data fields can be found on the FAQ page of World Happiness Report: https://worldhappiness.report/faq/

The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:

EDA (10%):

Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Merge and clean the data. Explain what you did.
- What are the central tendencies of happiness score over the years? Did they increase or decrease?
- Which countries have stable rankings over the years? Which countries improved their rankings?
- Visualize the relationship between happiness score and other features such as GDP, social support, freedom, etc.
- Find out what features contribute to happiness. If you are the president of a country, what would you do to make citizens happier?

(Note: EDA should be performed on the training sets only)

Modeling and Question Answering (10%):

The happiness rankings in the datasets are determined by happiness scores only. Now we want to predict the ranking using a machine learning approach. Build three models based on data from year 2015 to year 2018. Explain how each model works (briefly introduce the machine learning algorithms behind them). Predict the happiness ranking for the year 2019 (drop the "overall rank" and "score" columns first). Compare your rankings to the original rankings in "2019.csv". How does each model perform? Invent your own formula to calculate happiness score using features of your choice.

Project Report (10%):

Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):

Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file/notebook as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.

Choice 4: Can we predict whether a Hotel Booking will be cancelled?

Mentor: Saketh Chintapalli <schintapalli@cs.stonybrook.edu>

When it comes to hotel bookings, customers have a variety of options and deals and are sometimes often cancelling certain bookings for several reasons. Given hotel booking data for two major hotels, can we predict whether a customer will cancel the booking or not? We will explore the main concepts of EDA and modelling classification algorithms in this project.

Datasets:

The “Hotel_Bookings.zip” file contains the dataset to be used for this project and a file describing the various columns in the data. You must split the dataset yourself into training, testing, and cross validation data(when required).

The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:

EDA (10%):

Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Which country saw the most hotel bookings according to the data?
- What is the distribution like for both hotels with respect to price of a room per night?
- Which months are the most busy for both hotels? Which months see the most expensive per night costs?
- Which months see the most cancellations for both hotels?
- Examine distributions of bookings vs market segment.
- Which room type was most commonly booked? Most commonly cancelled?
- What percentage of the data recorded cancellations for each hotel?

(Note: EDA should be performed on the training set only)

Modeling and Question Answering (10%):

Apply three machine learning algorithms to predict whether or not a customer will cancel a booking. Train your models on the training set, and make predictions for the test set with the “is_canceled” and “reservation_status” columns dropped. Evaluate the accuracy of your predictions. If your predictions are not so accurate, what do you think is the reason? Use other evaluation metrics to evaluate your models (Precision, Recall, F-score). Split the data further to include a cross validation set. Did this improve your model’s performance on the test set?

Project Report (10%):

Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):

Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.

Choice 5: Can we predict the price range of a Mobile Phone based on its properties?

Mentor: Saketh Chintapalli <schintapalli@cs.stonybrook.edu>

Nowadays, a cellular phone has become an essential part of people's everyday lives(for better or worse). This has budded an increased level of competition between manufacturers to produce a phone in an optimal price range with the best features available to attract customers. This project will explore multi-class classification using various methods along with in-depth EDA.

Datasets:

The "Mobiles.zip" dataset will contain all the data you need for this project. It is your job to split this dataset into training and testing data. The data description is also available in the zip file. The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:

EDA (10%):

Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Examine the various columns with respect to price range and their distributions
- Compare and differentiate the distributions with respect to the different values of the target variable (price range).
- Try to perform clustering on the dataset into 4 clusters(without using the price range variable) using any clustering algorithm you know. Can you visualize the clusters in 2D or 3D using PCA?
- How did the clustering algorithm above work with respect to the price range labels? Differentiate using two different scatter plots(one color coded based on cluster label and the other based on price_range code).
- What features seem to be the most important ones? Perform a correlation analysis before your prediction task.

(Note: EDA should be performed on the training set only)

Modeling and Question Answering (10%):

Build three models, train them on the training set, and predict the outcome on the test set (after dropping the outcome column in the test set). Explain how each model works (briefly introduce the machine learning algorithms behind them). Evaluate the performance of each model based on the original outcome in the test set. For each model, visualize the confusion matrix. Which model works best? From the confusion matrix, what did you observe? Did certain models misclassify certain categories more than others?

Project Report (10%):
Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):
Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.

Choice 6: Can we predict the Selling Price of a House?

Mentor: Saketh Chintapalli <schintapalli@cs.stonybrook.edu>

There are many factors which affect the selling price of a property. This includes but is certainly not limited to the number of bedrooms, location of the property, crime rate in the area, etc. Given a dataset with various attributes of a house, we will learn to predict the selling price of that house

Datasets:
The “houses.zip” file on Blackboard contains data for all the houses sold in King County , Washington. Description of the data fields can also be found in the zip file on Blackboard.

The project has four components: exploratory data analysis, modeling and question answering, project report, and demo:

EDA (10%):

Get familiar with the dataset and decide what features and observations will be useful. Make good use of visualizations. Specific tasks may include but are not limited to:

- Visualize the correlation between all of the variables with Sale Price. Which variables seem to be the most positively correlated. Most Negatively Correlated?
- Examine the distribution of sales with respect to the month in which the house was sold. Do the same thing based on the year. Will it be useful to include these in separate columns? How correlated are these new columns with SalePrice?
- Examine the general distributions of the columns. Try creating some non-standard plots (box plots, area charts, stacked bar charts) to explain the columns and plots which tell interesting stories about the data.
- Try to create your own scoring function based on all the variables and compare how well your function did with respect to the "grade" column. Will your scoring function be useful to add as a column for your modelling(if you're doing this, make sure you don't use sale price as a variable in this function)?
- Can you create a visual map showing the most expensive areas of Kings County based on the data given?

(Note: EDA should be performed on the training sets only)

Modeling and Question Answering (10%):
Build three models based on data. Explain how each model works (briefly introduce the machine learning algorithms behind them). Predict the sale price of each house in your test set. How well do these three models work? Did you perform any feature engineering to make your model work better? What metrics did you use to evaluate your predictions?

Project Report (10%):
Write about your methods and findings in 2-3 pages. Include visualizations to prove your point. The report should be formatted like a research paper.

Demo (5%):
Sign up for a Zoom session with the mentor to present your project. Both team members should be present. Prepare to answer questions.

Submission:

Submit the following on Blackboard:

1. **Code** in a zip file as attachment
2. **Project report** in pdf format as attachment

You are also required to do a presentation on Zoom.