List of Final Projects

Course: Introduction to applied data science (PHYS247) Spring 2020
Instructor: Prof. Bahram Mobasher; TA: Nima Chartab Due Date: June 12, 11:59 p.m.

You need to choose a project topic by April 28. The suggested projects are listed below, if you want to choose a project which is not listed below please contact me at nima.chartab@email.ucr.edu. **To select a project, fill out the following google sheet: Click Here! Each project can be chosen by a maximum of 4 students. So, please check your classmates' choice on google sheet before choosing your own.** I encourage you to work in groups and share your thoughts and codes with your classmates who will choose the same project. However, grades will be attributed individually based on: 50% for the overall quality of your model and code, and 50% for the written report. For the former, you will share your .ipynb file as you do in your homework and for the latter, you need to write a short report (2-5 pages) which includes a brief description of the project and data set, data processing and cleaning procedure, plots and figures, explaining the method you used to build a model and evaluating your model's performance. Both of them are due on June 12, 11:59 p.m.

Project 1: COVID-19

In this project, you will use daily time series of COVID-19 including confirmed, deaths and recovered cases to fit SIR (susceptible-infected-removed) epidemic model parameters for different countries using MCMC (Markov chain Monte Carlo), and predict the pandemic evolution. Use any method you will learn in the course to gain insight into the impact of different factors in the transmission of COVID-19.

Dataset: https://github.com/CSSEGISandData/COVID-19

Project 2: Titanic

On April 15, 1912, the British ocean liner Titanic sank into the North Atlantic Ocean after colliding with an iceberg. In this project, you will use the passenger data (age, gender, etc.) to build a predictive model for the survival of the Titanic passengers. Apply any machine learning method you will learn in the class to predict which passengers survived the tragedy.

Dataset: https://www.kaggle.com/c/titanic/data

Project 3: House Prices

In this project, you will be given a data set with all aspects of residential homes in Ames, Iowa, and you will create a predictive model for the final price of each home. Apply any machine learning method you will learn in the class to predict the final price of homes in the test sample.

Dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

---

Project 4: Santander Customer Satisfaction

In this project, you will be given an anonymized dataset containing a large number of numeric variables from Santander Bank. In the training sample, there is "TARGET" column which equals one for unsatisfied customers and 0 for satisfied customers. Use any machine learning method you will learn in the course to identify satisfied/dissatisfied customers in the test sample.

Dataset: https://www.kaggle.com/c/santander-customer-satisfaction/data

---

Project 5: Restaurant Revenue Prediction

In this project, you will use data that is provided by TAB Food Investments (TFI) to predict revenue of the restaurant. The training sample includes the information of 137 restaurants and you will build a model based on the training sample and will apply it to 100000 restaurants in the test sample.

Dataset: https://www.kaggle.com/c/restaurant-revenue-prediction/data

Project 6: Credit Card Fraud Detection
In this project, you will use anonymized credit card transaction information to identify fraudulent transactions. Due to confidentiality issues, you will work with only numerical input variables which are the result of a PCA transformation. Use any machine learning method you will learn in the class to build a predictive model to classify transactions. .
Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud

Project 7: Leaf Classification
You will be given 1584 images of leaf specimens with three sets of features: a shape contiguous descriptor, an interior texture histogram, and a fine-scale margin histogram. Use any machine learning method you will learn in the class to build a classifier and apply it to the test sample.
Dataset: https://www.kaggle.com/c/leaf-classification/data

Project 8: New York City Taxi Fare Prediction
You will be given a data set of New York City Taxi fare amount with the information of pickup, drop-off, pickup time and etc. Use any machine learning method you will learn in the class to build a predictive model for the fare amount, then apply the model to the test sample.
Dataset: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

---