Detecting Spam Comments in Famous Youtube Videos

This idea of this project is to create a classifier model to classify Youtube video comments to spam/not spam. Given the video address, Comment ID, name of the commenter, date of the comment, and the content of the comment, we can create a binary text classifier that will give out the result of the prediction (spam or not) and the probability of prediction

The dataset for this project is UCI's http://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection The dataset contains 5 columns (4 predictors, 1 label) and 1,956 rows

The main predictor of dataset is the comment string, which is a natural language comment, that can be analyzed through basic ML techniques (like Naive Bayes) as well as advanced NLP techniques (like sentiment analysis). The biggest indicator of spam comment may be any website link mentioned in the comment. However, having only one main predictor and only about 2,000 comments, and looking at those comment strings, one could tell they could be too short and have too much variance in content between each other, so although as a human it may be easy to classify those comments, for a machine however, it may be a challenge to make a classifier with high accuracy regardless of the technique being used.

Predicting Appliances Energy Consumption Based on Ambient Conditions

The idea of this project is to create a regression model that would estimate energy consumption of home appliances based on temperature and humidity of the surroundings. Each row of the dataset contains multiple sets of temperatures and humidities over a period of time and the energy consumption is averaged over that time, So, there needs to be a regression model that would consider time-sensitive data

The dataset for this project is UCI's http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction It contains 29 columns and 19,375 rows

The dataset has all numeric values which makes it easier to plug them into any ML algorithm. Also, the dataset provides relatively high number of samples, but it only offers few features/indicators (temperature, humidity values) which may not tell the full story, and therefore, may not allow to generate a correlation/model of the energy consumption with those features. For example, from a human perspective, it is very hard to estimate energy consumption from other values. Also, by eyeballing the dataset, it has lots of missing values which may result in extra assumptions and other issues like over/underfitting

Predicting Gross Income of Movies

The idea here is to create a regression model that would put into account movie's release year, country, budget, director, genre, ratings and other features to predict movie's gross (revenue). The

dataset contains movies from 1986 to 2016 with features of mixed data types and requires additional pre-processing

The dataset for this project is Daniel Grijalva's Kaggle https://www.kaggle.com/danielgrijalvas/movies It contains 15 columns and 6820 rows

The dataset provides many different indicators to do the task with numerical, date, and string data types. Many string columns will need to encoded to make them useful in the regression task. The dimensionality of the dataset, relatively, makes a lot of sense although other features like movie critic reviews and others would've been helpful in estimating the gross. I think that many ML algorithms would do a good job with this dataset

Predicting the Cause of Wildfires

The idea in this project is to create a cause predictor/classifier of forest fires given region, location, date, and many other features. The huge SQLite dataset of 1.8M rows and 39 columns can be used to predict fire causes in different locations in the US and could used to extract many other insights.

The dataset for this project is Rachael Tatman's Kaggle https://www.kaggle.com/rtatman/188-million-us-wildfires/home

The dataset provides many feature columns with string, numerical, and date data types but not many of them are deterministic in the prediction task. Looking at columns like fire code, fire name, fire year, fire size, location, agency, and so on does not help me (as a human) predict fire causes. There needs other kind of background info like geographic location, topography of the area, vegetation, weather conditions, precipitation and so on. So it is highly doubted that any ML model would get any great accuracy classifying fires' causes into their respective classes.