Daniel Weissberger
7/14/2018

Capstone 1 – Ideas

1) **Predicting high risk credit card customers -** UCI has a database which provides data on credit card customers. It contains information on their marital status, gender, education, age, credit limit and payment history (up to the last 6 months before the default).This information could be used to determine which customers are at highest risk of defaulting. Credit card companies could potentially use this information for case management purposes or alternatively when assigning a credit limit to a customer.

The data for this dataset is in the form of a 30,000 row csv file and it is all numerical data:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Credit Card Defaults Dataset

2) **Predict how much a movie will generate in revenue –** This is a database of movies which contains information such as budget, release date, genre, runtime, production company/country. The challenge would be to use regression to predict gross revenue of a film based on some of these features. This information is highly valuable since it would allow production companies to look at predicted revenue and then make adjustments to certain features and budget accordingly.

The dataset contains 4800 records and is a combination of numerical, text and object (json) format. The columns/fields are listed below:

Budget, genres, homepage, id, keywords, original_language, original_title, overview, popularity, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, vote_average, vote_count

Movie Dataset

3) **Exploratory data analysis of NYC parking tickets** – This is a dataset of parking tickets issued in the NYC area. It contains data such as Vehicle make, Violation Location, Time of Citation, Feet from Curb, Hydrant Violation, etc. This could be very useful to NYC parking enforcement since it would essentially establish a heat map of parking tickets so they could identify deficits in parking attendant coverage

This dataset contains 105,000 rows of parking violation data from 2017 and the following columns of mixed data type (integer, text):

Summons Number, Plate ID, Registration State, Plate Type, Issue Date, Violation Code, Vehicle Body Type, Vehicle Make, Issuing Agency, Street Code1, Street Code2, Street Code3, Vehicle Expiration Date, Violation Location, Violation Precinct, Issuer Precinct, Issuer Code, Issuer Command, Issuer Squad, Violation Time, Time First Observed, Violation County, Violation In Front Of Or Opposite, House Number, Street Name, Intersecting Street, Date First Observed, Law Section, Sub Division, Violation Legal Code, Days Parking In Effect, From Hours In Effect, To Hours In Effect, Vehicle Color..etc

Parking Tickets in NYC

4) **Predict housing prices-** This dataset is from a Kaggle competition and contains 79 variables detailing almost every feature of residential homes in Ames, Iowa. The goal of the competition is to predict the sales price of homes in the test dataset using regression and other machine learning techniques. This information would be valuable to real estate agencies looking to establish a sales price for a home or recommend an offering price to a buyer.

This dataset contains about 1459 rows of data including the following features: Sales price, Building Class, Zoning, Lot Frontage, Lot Area, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Year Built.. Etc.
House Prices Ames, Iowa