Capstone Project Ideas

Predicitng Medical Appointment No-Shows

Appointment no-shows are a significant problem in the medical industry. By calculating the probability of a pationt no-show, medical facilities can optimize their bookings to reduce the costs of under and over-booking.

Can patient information such as age and health problems, as well as time of day and day of week, be used to predict the probability of them making their appointments, and provide information that can be used to create a smarter booking system that can reduce the problems associated with more or fewer patients showing up than expected?

The dataset can be found on Kaggle at the following link: https://www.kaggle.com/joniarroba/noshowappointments. The data consists of ~110,000 medical appointments each with 15 variables, including appointment day, day scheduled, age, neighborhood, show or no-show, and patient medical conditions such as hypertension, diabetes, alcoholism, and handicap.

Bikeshare - Predicting Usage and Availability

Bikesharing services allow customers to rent and return bicycles at stations throughout a city. There are two potential issues that can cause major inconveniences for the users. If a user expects to drop off their bicycle, but the station is full, or if no bicycles are available when the user wants one.

There many questions to ask of the data. Can the total ridership be predicted based on weather conditions, day of week, holiday, for example. How commonly do stations run empty or fill up, and which stations does this affect? For a station affected by shortages of bicycles or return space, can the availability of a bike or return space be predicted?

The bikesharing data is found at https://www.kaggle.com/benhamner/sf-bay-area-bike-share. This set contains 4 files: station, status, trip, and weather:

station: small table that shows name, latitude, longitude, dock count, city, and install date for each of the 70 stations.

status: List one minute status intervals showing number of bikes and docks available for each of the 70 stations.

trip: each trip in a two-year period is an obervation, the variables include: start time, end time, start station, end station

weather: daily weather obersvations including temperature, humidity, wind speed, and precipitation.

Predicting Income

There are many uses for determining income level in marketing. This project would explore using demographic variables to predict income class.

The data can be found at https://archive.ics.uci.edu/ml/datasets/census+income. The data consists of ~48,000 obeservations of 15 independent variables including age, gender, education, marital status, occupation, hours worked per week, and native country, as well as the dependent variable, which is whether the income level is above $50,000 or not.