Final Project Ideas

BRANNON WALDEN

Classifying Building Permit Use Types for the City of Montgomery

Business Understanding: City employees must manually enter the classification of each building permit into a computer system. A program that would automatically classify the [Use Type] as “Commercial” or “Residential” based on data already being entered by the user (e.g. address, business name, project synopsis) could lead to fewer input errors and allow employees to concentrate on other tasks saving the city time and money.

Data Understanding: Building permit data is publicly available via a Socrata open data platform and contains records from 2014 to present. There are 17.5K rows, and 35 columns. Running summary statistics, histograms, correlation coefficients, and other analyses will help identify key features and help to align data understanding with the business understanding.

- https://data.montgomeryal.gov/Permits/Building-Permit-2014-Present-Download-/qvzc-ejq2

Data Preparation: File is generally clean but likely includes some missing values, outliers, and unnecessary fields. Identify key features for the model.

Modeling: Potential candidates for modelling include Logistic Regression, Decision Tree, or Support Vector Machines. May also consider a multi-label classifier for “permit type” instead.

Complexity Level: Low/Med

Predicting Emergency Levels of Calls for Service for the San Francisco Fire Department

Business Understanding: San Francisco has a large call for service volume, but many calls are not life-threatening. However, incidents that were serious included public and/or rescuer fatalities, injuries, and/or property damage. If first responders had an accurate rating system that predicted the level of emergency based on incoming call data, dispatchers may choose different levels of personnel and/or apparatus to send to the call location.

Data Understanding: Call data is publicly available via Socrata's open data platform and contains records from 2000 to present. There are 4.64M rows, and 34 columns. Running summary statistics, histograms, correlation coefficients, and other analyses will help identify key features for modelling. A clustering algorithm may also be helpful to identify the best class candidates. Must determine if there are enough records for each classification to make accurate predictions.

- https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3

Data Preparation: Data needs to be classified based on multiple features. Number of public and/or rescuer fatalities and/or damage costs could be used to create three labels of emergency classification (high, med, low). Missing values and outliers may need to be dropped. Standardization and/or dimensionality reduction should also be considered to reduce model error, effects of outliers and computational costs.

Modeling: Potential candidates for a multi-class model include Linear Discriminant Analysis, Naïve Bayes, K-Nearest Neighbors, etc. Caution: There could be too few records with a "high" classification, for example, to accurately make predictions for high level emergencies. The error rate for such modelling would need to be investigated using different techniques.

Complexity Level: Med/High

Identifying Threats from GAIA Spacecraft Observations

Business Understanding: In addition to star data, the GAIA observatory recorded many objects within the solar system that could be potential dangers to Earth. To protect earth, determine which objects are heading towards us that may pose a future threat, or that have a particular interesting characteristic yet to be determined.

Data Understanding: The object data contains a sample of 14,099 SSOs for a total 1,977,702 different observations. Summary statistics, histograms, correlation coefficients and other analyses will be useful to determine key features. Dimensionality reduction or clustering may be employed to help identify classifications of objects.

- http://cdn.gea.esac.esa.int/Gaia/gdr2/sso_observation/csv/

Data Preparation: SSO observations are contained in four separate CSV files which will need importing and joining. Caution: the complexity of the data and the math to determine orbital paths could be out of reach to the novice. Standardization and dimensionality reduction are likely necessary.

Modeling: Potential candidates for modelling may include Support Vector Machines, K-Nearest Neighbors, Random Forests or even Neural Networks. Caution: The very large volume of data under consideration may lead to high computational costs in which sampling or a cloud computing environment may be required.

Complexity Level: Very High