**XGBoost for Regression and Classification: Airbnb Price Prediction and Forest Cover Types**

Matthew Brigham and Matthias Quinn

STA 685

Final Project

Fall 2021

**What is XGBoost and how will it be used?**

XGBoost is a specific algorithm for gradient boosted trees. Gradient boosting is typically a very slow process; what makes XGBoost special is its speed and performance compared to other gradient boosting algorithms. This specialty makes it particularly useful for large datasets, in addition to usually having higher model accuracy than other types of supervised methods. We plan on showing how XGBoost can have superior performance on large and small datasets in the context of both regression and classification. We will be working on two separate applications: Airbnb Price Prediction (regression) and Forest Cover Type (classification).

**XGBoost and the Airbnb Application**

The Airbnb Price Prediction application is a regression problem where we will try to predict the price of an Airbnb listing. The company “InsideAirbnb” is an activist project that scrapes data from Airbnb and stores it on their website for up to a year. The purpose of InsideAirbnb is to understand how Airbnb is being used to compete with the residential housing market. They scrape data from Airbnb for cities all around the world, but we chose a dataset from a single city: Austin, TX. The data was scraped on October 14, 2021 and represents all active listings at that time. There were 10809 active listings and there are 74 features for each listing. The data can be accessed online at [1]. Table 1 shows a description of *some* of the variables that describe each listing. **XGBoost will be used in a regression problem on the 74 variables to predict the price of an Airbnb listing. Some of the variables are text data, so encoding *may* be used.** One reference for this application is Alin-Cristian Preda, 2020. [2] Their paper used XGBoost to predict nightly prices of Airbnb’s for many cities.

|  |  |
| --- | --- |
| Name | Text string name of the property given by the owners. |
| Description | Text data string describing the property created by the owners. |
| Host is Superhost | A binary variable that is True if the host is a super host, else False. |
| Property type | A text description of the property, e.g., “Entire residential home”, “Private room in the residential home”, “Entire guesthouse”, etc. |
| Bathrooms | The number of bathrooms. |
| Bedrooms | The number of bedrooms. |
| Amenities | A list of strings describing the amenities the property has, e.g., “Free street parking”, “Air conditioning”, “Patio or balcony”, “Wi-Fi”, etc. |
| Price | Price of the unit on the date the data was scraped. |

Table 1: A sample of the 74 variables in the data set with descriptions.

**XGBoost and the Forest Cover Type Application**

The Forest Cover Type application is a classification problem where we will try to predict the type of forest cover (tree species) for a given patch of land. The Forest Cover Type dataset was donated to the UCI Machine Learning Repository in 1998 as is freely available on their website. It has 54 variables and 581,012 observations/patches of land. Each observation is a 30mx30m patch of land. Table 2 shows a sample of *some* of the variables in the dataset. The data can be accessed at reference [3]. Sameer et al [4] successfully implemented XGBoost to classify the Forest Cover Type. There are 7 possible forest cover types: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas Fir, and Krummholz. **We will attempt to use XGBoost for the classification of forest cover types using the 54 variables in the dataset and any additionally engineered features.**

|  |  |
| --- | --- |
| Elevation | Elevation of patch in meters |
| Slope | Slope of patch in degrees |
| Horizontal distance to hydrology | Horizontal distance from patch to nearest surface water |
| Vertical distance to hydrology | Vertical distance from patch to nearest surface water |
| Soil Type | There are 40 columns of specific soils. These columns are binary: either 1 or 0 depending on if that soil is present in the patch. |
| Cover Type | This is the response we are trying to predict. This is a categorical variable with 7 possibilities, each integer corresponding to a type of tree cover for the patch. |

Table 2: A sample of some of the 54 variables with descriptions.

References:

[1] <http://insideairbnb.com/get-the-data.html>

[2] <https://research-api.cbs.dk/ws/portalfiles/portal/62184289/898508_Dissertation_Alin_Preda.pdf>

[3] <https://archive.ics.uci.edu/ml/datasets/covertype>

[4] <https://www.researchgate.net/publication/353675864_Rapid_Forest_Cover_Detection_Using_Ensemble_Learning>