Final Project Proposal
Machine Learning Fall 2021
Sara Clark & Christine Pourheydarian & Jiawei Liu

Overview

- A description of your problem and motivations.
    - Problem: Wellesley students are familiar with Boston by senior year. Yet many students end up moving to other large cities when they graduate in order to start new jobs. Many move to cities that they are not fully familiar with, and so need to do research to see what neighborhoods they should move to. This research can get overwhelming, since it often involves reading individual people's opinions online.
    - Motivations: We found the Airbnb data helpful for making such decisions, since there are comprehensive reviews / other dimensions of data for neighborhoods of each house listing. We are motivated to use this expansive data in order to see if we can create a more accurate way for individuals to determine what neighborhoods of cities they would like best, through the lens of a data set that is often not used for such purposes we hope to give such individuals new, but useful information to guide their decisions.
- Existing Work
    - We are anticipating creating our model using Airbnb data. Since Airbnb is a company that already uses machine learning to shape its product, there are many resources and articles discussing their usage of machine learning. Additionally, the widely available data encourages other individuals to create their own algorithms based on the data, which allows users to create models that do everything from predict prices to make recommendations.
    - Predicting Airbnb prices with machine learning and deep learning
    - How to build a Recommender System for Airbnb in Python
    - Airbnb – using AI to evaluate if a guest is trustworthy - Digital Innovation and Transformation
    - How Airbnb uses Machine Learning to Detect Host Preferences

The Data

- Our data is broken down into two categories based on the two stages of our project.
- Neighborhood Classifier/recommender based on Airbnb Similarity
- Stage One: Build Boston Neighborhoods Classifier

    - Use numerical data and text from reviews, neighborhood descriptions and home descriptions to accurately classify what neighborhood an Airbnb is located in.
    - Word cloud: find out the top several feature/words that are the most related with each neighborhood
    - Will use Boston Airbnb Open Data.
    - Note: Some entries are not linked to a specific neighborhood, so we will need to filter the data set to include only data that is labeled with a neighborhood.
- Stage Two: Cross-city application
    - Use the trained model on one city to predict on neighborhood descriptions on another city, in order to find neighborhood similarity between cities
    - Inside Airbnb. Adding data to the debate.
        - Note: This data set has the same features as the Boston Airbnb dataset. This would mean that we are able apply our algorithm onto other cities by simply changing the input data, assuming we did not hardcode values specific to Boston.
- Additional Applications (if we have extra time, we will do this)
    - Applying the model to non-Airbnb source
        - Input text from Wikipedia, Yelp, and other review sources. This will not be used to modify but rather to test the algorithm (how well does the algorithm perform on other descriptions of neighborhoods).
        - The dataset that you will be using, with a link if relevant.
- Featurization:
    - 92 features, ranging from full sentences to single words to integers. We will likely use less features in order to prevent overfitting (applying dimensionality reduction).
    - We will scale our numerical data and convert non-numerical values (eg yes/no) to numerical data.
    - We anticipate having to separate our approach to processing data that can be processed numerically with blocks of text, and will potentially approach these sets of features with different algorithms but ultimately combine the results of the models.
    - We will see what the weights are of different features and will use this information accordingly, especially for features with discrete values.
- Splitting Data
    - We will first randomly permute the rows of the dataset, then split into training and testing (75% training, 25% testing) and we plan on splitting our testing data using cross-validation.

        - *We are concerned whether randomly shuffling the data will ensure we have all neighborhoods in both training and testing data. Should we split training and testing data so that each neighborhood is equally represented in the testing data instead?*

Data Analysis

- Classification Algorithms
    - Bag of words: tfid vectorization
    - kNN
    - Decision Trees
    - SVM
    - Logistic Regression
    - Neural Network (or Perceptron)
    - We want to use a wide variety of classification algorithms so that we can take advantage of the ensemble method to classify our data in order to produce more accurate results. We are operating under the assumption that we are able to use sklearn, which would allow us more capacity to focus on other details of the project rather than creating the algorithms from scratch.
- Evaluation
    - Due to the fact that we are not classifying based off of a binary but rather a variety of neighborhoods, we will rely on accuracy to evaluate our results.
    - Will you consider any tradeoffs between accuracy, interpretability, and fairness of your algorithms?
    - We believe that our results will have high interpretability since we are using supervised learning, and our primary aim is to produce an accurate classifier. However, the host's descriptions of their Airbnb's neighborhood will likely be positively skewed. In order to get a more fair and accurate representation, we will attempt to incorporate reviews of Airbnb's to get a more balanced view of a neighborhood. In any area, particularly diverse cities like Boston, some neighborhoods carry a more negative connotation and we are aware that some reviews may reflect that, but we hope that the host's more positive perspectives will balance out potential biases reflected in reviewers.
- Is the primary purpose of your task prediction, or do you also want to explain something about the data, e.g., analyzing features that are predictive of a class?
    - Prediction is a primary goal, but we also want to get insights about what different neighborhoods are like (i.e. chill, good vibes, boring, etc.), and use these insights as guides for future neighborhood choice/recommendations

Project Details

- What aspects of your project do you anticipate will challenge you most?
    - Processing the data
        - Removing features (deciding # of features and which to use, hyperparameter tuning)
        - Converting non-numerical features to numerical
        - Handling rows with missing features
        - Filtering data to only include rows that specify a given neighborhood
    - Creating an accurate Boston Neighborhood classifier/ recommender.
- Project Outline and Responsibilities:
    - Sara- tasks in orange
    - Jiawei - tasks in purple
    - Christine - tasks in green
    - Stage One:
        - Process data
            - (1)Removing rows without neighborhoods
            - (2)Converting non-numerical to numerical
                - Potentially removing unnecessary features
            - (2)Feature scaling
                - We are not sure how to approach this step given the wide variety of feature types. Any advice would be much appreciated!
            - (2)Text Analysis
                - Remove neighborhood names from text-based feature columns used for training
                - Bag of words: tfid vectorization
                - Link reviews to correct neighborhood
            - (3) Randomly permute rows
            - (4) Split into training, testing and prepare for cross-validation
        - Create classification models
            - kNN
            - Decision Trees
            - SVM
            - Logistic Regression
            - Neural Network (or Perceptron)
            - *Should numerical data and text features be classified together or separate?*
        - Test and evaluate
            - Determine accuracy of model on testing data
            - Extract key data related to each neighborhood (e.g. most important words ... bag of words)
    - Stage Two:

        - Compare other city's neighborhoods to Boston
            - Change testing data to be other Airbnb data for other cities
        - Apply neighborhood classifier to a different data set
            - Change training data for the same code in Stage 1
                - input city name
                - pull corresponding file from kaggle and read it in?
    - Milestone 2: Presentation: everyone; will break down later on
    - Milestone 3: Paper writing: everyone; will break down later on