

# **Introduction to Machine Learning – Project Proposal**

## **RENTAL LISTING INQUIRIES**

### **GROUP MEMBERS**

1. Akhauri Prateek Shekhar
2. Swapnil Sagar
3. Akash Dobaria
4. Khushbu Durge
5. Urvi Patel
6. Anamika Paul

### **MOTIVATION**

Searching for the perfect home can be tiring; gathering all the details and browsing through endless listings. Structuring and making sense of all the available real estate data programmatically is difficult. To regularize and ease the same, the project will predict how popular an apartment rental listing is based on the listing content like text description, photos of the apartments, number of bedrooms, price, etc. Doing so will help renthop.com handle fraud control, identify potential listing quality issues, and allow owners and agents to better understand renters' needs and preferences.

### **AIM**

The objective here is to predict the number of inquiries a new listing receives based on the listing's creation date and features of the apartments/houses.

### **PROJECT PLAN**

To do the task, first we need to acquire data for the apartment rental inquiries from Internet. Then we should determine the attributes we want to use in the machine learning algorithm. At last we should find the best machine learning approach to give us the results.

#### **1. Data Acquisition**

Source: <https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries>

Data Source: [www.renthop.com](http://www.renthop.com)

#### **2. Feature Determination**

A huge number of features may affect the search for an apartment. Features like number of bedrooms, number of bathrooms, area, locality etc are some of the many features which would contribute for the prediction of the inquiries. As we make progress in the project we can add on other features required for it and which feature contributes to which extend towards determining the objective.

#### **3. Machine Learning Approach**

Initially, we will use the basic machine learning approaches learnt in class like Naïve Bayes, Random Forests, Multinoulli Logistic Regression, Decision Trees and Support Vector Machines in our project and determine the prediction output. For testing, validation, and evaluation purpose we plan to use 10-Fold Cross Validation, Precision, Recall and F1-Score. Depending upon the error rate values we may further make use of other methods of classification and prediction as well to get the best classifier.

### **EXPECTED RESULT**

The variable – **interest\_level** will have an output in the form of three classes – **High, Medium and Low** and the output file will be in the form of an excel sheet with the Listing Ids of the apartments and their respective interest level probabilities of high, medium and low.