**Synopsis**

**HOUSE PRICE FORECASTING AND RECOMMENDATION SYSTEM**

**( Machine Learning project)**

# INTRODUCTION:

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors.

Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry. Notably.

In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, in

it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average.

This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in Calgary between 2017-2020 based on data . The aim of our project was to build a predictive model for change in house prices in the year 2022 based on certain time and geography dependent variables.

**BACKGROUND:**

The field of Data Science is rather young, having taken form over the last halfcentury as a discipline distinct from statistics. It is also rapidly growing with many interesting advancements in recent years, most notably within Machine Learning (ML). This has resulted in an increase in media attention as well as funding of AI related businesses and research projects. In 50 years of Data Science [1] Donoho comments on the history of Data Science and questions whether it is really different from statistics. With regards to Machine Learning, he points to a study he conducted that compared a set of highly-cited and glamorous classifier methods such as Random Forests and k-Nearest neighbour to a simple linear classifier applied on the same problem. The study found that the simpler method did not only perform similarly, but had a lower worst-case regret. This suggests that when benchmarked, more advanced ML algorithms are not necessarily better when put in practice, which highlights the need for making algorithm comparisons. A study using similar techniques was made on predicting the sales price of used cars. [2] This problem is similar to predicting house prices and arguably simpler because it is dealing with cars, commodities that aren’t geographically fixed and are often highly standardized. The methods used were Multiple Linear Regression, k-Nearest Neighbours, Naïve Bayes and Decision Trees, including Random Forest. Cross validation was used for finding the optimal hyperparamaters, such as the ’k’ for k-NN. The house pricing problem was approached by Baldominos et al. [3] from the viewpoint of finding investment opportunities. They formulated the regression problem and used several Machine Learning algorithms such as k-Nearest neighbour, variations of neural networks and decision trees. Another study by Oxenstierna [4] investigated it for the purposes of valuation of houses. The 2 CHAPTER 2. BACKGROUND 3 data set included 5000 entries. Again, the k-Nearest neighbour method was used as well as Artificial Neural Networks, to minimize the median absolute percentage error of the prediction. The methods performed similarly at around 8-9 % Median Absolute Percentage Error. 2.1 Machine learning algorithms

# Objectives & Data:

The project goal is to predict sale prices for homes from a given Data by analyzing and

understanding patterns in data

Data:- Our training data consists of 1,460 examples of houses with 79 features describing every aspect

of the house. We are given sale prices (labels) for each house. The training data is what we will

use to “teach” our models.

****Testing****: The test data set consists of 1,459 examples with the same number of features as the training

data. Our test data set excludes the sale price because this is what we are trying to predict. Once our

models have been built we will run the best one the test data and Deploy The project to Heeroku

****Task****: Machine learning tasks are usually split into three categories; supervised, unsupervised and

reinforcement. For this project, our task is supervised learning(Regression and classification Analysis).

*S*upervised learning uses examples and labels to find patterns in data

It’s easy to recognise the type of machine learning task in front of you from the data you have and your

objective. We’ve been given housing data consisting of features and labels, and we’re tasked with

predicting the labels for houses outside of our training data.

**SCOPE:**

Having lived in India for so many years if there is one thing that I had been taking for granted,

it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing

prices have recovered remarkably well, especially in major housing markets. However, in the

4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in

the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first

time there was a decline since Q1 of 2017. The decline has been partly attributed to political

uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency

among customers and also the comparison can be made easy through this model. If customer

finds the price of house at some given website higher than the price predicted by the model, so

he can reject that house.

# REQUIREMENT ANALYSIS:

* **SOFTWARE REQUIREMENTS:**

**Operating System: Windows 10**

**Front end: HTML,PHP,Python**

**Back End: MS Workbench**

**LIMITATIONS:**

The local data will be requested from the Svensk mäklarstatistik [3]. The request contains a list of

features, that matches the public dataset's features, that is desired to be available when the data is

sent. There is no guarantee that the data will be available in time nor contains the exact requested list of

features. Thus, there might be a risk that the access will be denied or delayed. If so, the study will be

accomplished based only on the public dataset. Moreover, this study will not cover all regression

algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques

to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide

area and several training methods that do not fit in this study.

**SURVEY OF TECHNOLOGIES:**

the “HOUSE PRICE FORECASTING AND RECOMMENDATION SYSTEM” will be used to for

predicting the house price, forecasting that price and also to get best recommendation according to

users requirement. This application can be run using website.

We are using Python3 for making machine learning model and Python flask for connectivity and HTML

to develop our web page.

# We are using anaconda which contains a software Spyder and Jupyter Notebook. Spyder contains all updated

# and latest libraries of python which will be very useful for implementing machine learning model linear

# regression, ARIMA model and content based Recommendation system. Sublime Text 3 will be used for

# implementing HTML web page which will be user interface

**REFERENCES:**

* Data source, https://raw.githubusercontent.com/shughestr/PIMS\_2020\_Real\_Estate\_ data/master/sample\_clean.csv.
* Kaggle competition, [https://www.kaggle.com/c/house-](http://www.kaggle.com/c/house-prices-advanced-regression-techniques)prices-[advanced-regression-techniques](http://www.kaggle.com/c/house-prices-advanced-regression-techniques).
* Zillow zestimate, [https://www.zillow.com/blog/zestimate-](http://www.zillow.com/blog/zestimate-updates-230614/)updates-[230614/.](http://www.zillow.com/blog/zestimate-updates-230614/)
* Kam C Chan, Patric H Hendershott, and Anthony B Sanders, Risk and return on real estate:

evidence from equity reits, Real Estate Economics

* B. de Ville, Decision trees, Wiley Interdisciplinary Reviews: Computational Statistics **5** (2013),

no. 6, 448–455.

* T. G. Dietterich, An experimental comparison of three methods for constructing ensembles

of decision trees: Bagging, boosting, and randomization, Machine learning **40** (2000),

* Franz Fuerst and George Matysiak, Analysing the performance of nonlisted real estate funds: a panel

data analysis, Applied Economics **45** (2013), no. 14, 1777–1788.

* Richard J Herring and Susan M Wachter, Real estate booms and banking busts: An international perspective, The Wharton School Research Paper (1999), no. 99-27.
* H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of educational psychology **24** (1933), no. 6, 417.
* R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society:

# Project Developed By:

**MR. Barkade Rahul Balu.**

# Project Guide:

**Prof. Anil Gaikwad**

**THANK YOU !!!**