* + **Introduce the problem**

My topic of research will be "House Price Prediction in chicago,IL". Since housing market is in the uptrend since last few years and with so many macro economic changes, it's difficult to understand the trend in future and i would like to address this issue using Data science and predict the housing price movement. The goal was to predict the price of a given apartment according to the market prices taking into account different “features” that will be developed in the following sections. Lot of different questions can be answered using the model . for example, how is the interest rate change impacts the house price how the first time buyers impacting the house price change. The impact of job movement from major cities on housing price. the impact working pattern of professional on housing market.

* + **Justify why it is important/useful to solve this problem**

The reason that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities. We can use Machine Learning in Finance, Medicine, almost everywhere. That’s why I decided to conduct my project around the Machine Learning.

* + **How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?**

Predicting something to near time accuracy is always challenging but equally rewarding as well. And predicting house market trend is definitely rewarding if we can use it properly . The decision making becomes much easier when machines learn the trends by themselves and gives the trend based on different factors.

Without this solution in place, its always kind of guess work on predicting the market price move based on different indicators and that is why we need to have some system which can be used to understand and take a calculative risk.

* + **Explain where you obtained your data**

My data is mainly from keggle and I used the data source for Chicago housing prices in last few years. The main element in machine learning task for which a particular attention should be clearly taken is the data.the results will be highly influenced by the data based on where did we find them, how are they formatted, are they consistent, is there any outlier and so on. At this step, many questions should be answered in order to guarantee that the learning algorithm will be efficient and accurate.

* Organized and detailed summary of Milestones 1-3
  + EDA; include any visuals you think are important to your project

Milestone 1:

A lot of steps are taken to get, clean and transform the data. I am going to explain each one of them to show how they have been applied on my project

The data is broken down according to the number of bedrooms that are inside of the housing unit.

In Milestone 1, I mainly tried to find out the relation between different factors like interest rate, age of the building , total living area, highway distance and structure quality on the sale price of the house. Here are some of the visuals.

Scattered plot for Structure quality and sale price

Chart

Description automatically generated

Scattered plot for highway distance and sale price

Chart, scatter chart

Description automatically generated

Scattered plot for total living area and sale price

Chart, scatter chart

Description automatically generated

Scattered plot for total Age of the building and sale price

Chart, scatter chart

Description automatically generated

Data preparation

Cleaning is the first thing I did before starting the work. There were many attributes which are just for information purpose only and do not impact the sale price of the building. Like below example where parcel number do not have any impact on the sale price.

So there were many such attributes which needed to be dropped from the dataset before starting the model. All these attributes were dropped in this exercise. 'PARCELNO','LATITUDE','LONGITUDE','avno60plus'.

Chart, scatter chart

Description automatically generated

After dropping these attributes, the next step was to replace any null fields with the mean and that’s what I did , so that there cannot be any null value in any of the field.

Dropping duplicate rows was the next task in data preparation and filtering any outliers in the dataset. For my dataset I used below parameter to filter the outliers.

q\_low = df.quantile(0.10)

q\_hi = df.quantile(0.90)

df\_filtered = df[(df < q\_hi) & (df > q\_low)]

Once its done, I made sure there is no null or any duplicate in my dataset. And that’s how I prepared my Data for modelling.

**Model building and evaluation**

I applied a sampling technique in order to divide it into different subset having each its own utility. It is commonly assumed that more we have data to build a model more it will have tend to give good results. Usually the dataset is divided as follow with their respective utility :

80% -🡪 for training

20% 🡪 for test

For model building, I used LinearRegression and below are the details.

| **Coeff** |
| --- |
| **LND\_SQFOOT** | 2.423592 |
| **TOT\_LVG\_AREA** | 204.348508 |
| **SPEC\_FEAT\_VAL** | 2.933498 |
| **RAIL\_DIST** | 4.915724 |
| **OCEAN\_DIST** | -4.825054 |
| **WATER\_DIST** | 0.650888 |
| **CNTR\_DIST** | -3.360307 |
| **SUBCNTR\_DI** | 0.260530 |
| **HWY\_DIST** | 4.451223 |
| **age** | -1819.403469 |
| **month\_sold** | 404.997846 |
| **structure\_quality** | 62306.732606 |

Based on the model, the accuracy is around 68%.

**Conclusion**

I evaluated and compared each model to determine the one with highest performance. I also looked at how some models rank the features according to their importance. I also followed the data science process starting with getting the data, then cleaning and preprocessing the data, followed by exploring the data and building models, then evaluating the results and communicating them with visualizations.

As a recommendation, I advise to use this model by people who want to buy a house in the area covered by the dataset to have an idea about the actual price. The model can be used also with datasets that cover different cities and areas provided that they contain the same features. I also suggest that people take into consideration the features that were deemed as most important as seen in the previous section; this might help them estimate the house price better.