House Sales Recommendation System Planning

House Sales Recommendation System

This project is a recommendation system for Real Estate companies based on insights from exploratory data analysis of house sales data in King County, Washington, USA.

1. Define the problem
    a. Business Understanding

        Domain: Real Estate.

        Business: Buy and sell houses in the Real Estate market.

    b. Business Problems (presented by business experts)
        i. Which houses should be bought and for what price?
        ii. Once its bought when it's the best time period to sell it and for what price?
        iii. To rise the housing selling price, the company should do a renovation. So what would be renewal changes?

    c. Objective in business terms.

        Improve business profit by buying houses for a price value and sell them for a higher price.

2. Preliminary Solution Proposal
    a. Final deliverables
        i. Interactive and web accessible dashboard with reports and houses portfolio overview.

    b. Expected tools
        i. Python 3.8.0
        ii. Jupyter notebook
        iii. Streamlit
        iv. Heroku

    c. Tasks guidelines
        i. Do classic Data Analysis (Get to know the data).
        ii. Make Business Data assumptions.

iii. Rise 10 hypotheses from visualizing the Data Profiling and from intuitive Data Interpretation.

iv. Do visualizations to test hypotheses and some Feature Engineering if needed to prove hypotheses.

v. Formulate House recommendations to answer business team.

vi. streamlit main page of statistics and maps visualizations with interactivity for the CEO.

vii. streamlit reports page with first report of insights 27/10/21, with the principal hypotheses and mainly the ones that are translated into business actions.

viii. streamlit report page with houses recommendations to answer business team problems.

ix. make the dashboard of streamlit web accessible on heroku.

3. Data Collection

   Houses sales data from King County, Washington, between May 2014 and May 2015 gathered from a public source.

   Source: https://www.kaggle.com/harlfoxem/housesalesprediction

4. Data Cleaning

   Not needed

5. Data Integration

   Not needed

6. Data Analysis

   a. Create a copy of the data for exploration (sampling it down to a manageable size if necessary).

   b. Through Exploratory Data Analysis:

      i. Get:

         1. Data Dimensionality
         2. Attributes Name and Description
         3. Attributes Type (categorical, int/float, bounded/unbounded, text, structured, etc.)
         4. Convert data type (if necessary)
         5. Attributes Descriptive Statistics
         6. Identify missing values and Duplicated records

7. Data Granularity
    a. On geographic data
    b. On temporal data
    c. On categorical data hierarchy or taxonomies
    d. On numerical data, its discretization level or composition
8. Data Distribution
    a. Type of distribution (Gaussian, uniform, logarithmic, etc.)
9. Data Sparsity
    a. Outlier Analysis
    b. Correlation Analysis
        i. It can be done for numeric x numeric attributes, numeric x categorical, categorical x categorical.
        ii. Then calculate correlation between numeric x target, categorical x target.
        iii. And then assume which features may be relevant and which are redundant.
10. Attributes Usefulness for the task
11. Identify Data Inconsistency
12. Define some Business Data Assumptions
    a. Ver references Pedro
    b. Find some record anomalies, and some duplicates
        i. Houses with 0 bedrooms and 2 bathrooms
        ii. House with 33 bedrooms
    c. There is some difference between sub-regions in King County, so there must be a separation between regions, and have a specific analysis for each. Since a generalized analysis would nullify the correlation of each region, and would be harder to take insights. This region segmentation can be made in different granularities, that is it can be divided by North, South, and East Mountains, but it can also have a finer granularity that is for zipcode regions. Another finer granularity can be the street or street block by collecting that data from the latitude and longitude, and counting the number of

houses per region to assess if there is enough data for each street to get significant statistical data.

13. Verify the assumptions with business experts

ii. Find insights to solve the initial business problems, by following the preliminary tasks guidelines.

1. If it was a comparison question about a conjecture or an hypothesis (i.e. a statement) between different entities or sub groups, use summary statistics and other statistical tests to ensure that it is verified.
2. If it was a well defined straightforward question about a specific sub-group behavior or variable value, it is resumed to data queries and search tasks.

iii. Find new insights for the business team by setting a set of hypotheses.

1. Create a Mind map with the hypotheses
2. Hypotheses:

   H1: Houses that have a water view, are 20% more expensive on average.

   H2: Houses with year built older than 1955, are 50% cheaper on average.

   H3: Houses without basement are 40% bigger than with basement.

   H4: The growth rate of the houses price YoY (Year over Year) is 10%.

   H5: Houses with 3 bathrooms have a MoM (Month over Month) growth of 15%.

   H6:

c. Explore the data attributes that are good to be used by a regression model or other type of ML problem, depending on the task to do after interpreting insights, that could be predictions, classifications, segmentations.

d. Study how you would solve the problem manually.

e. Identify extra data or some feature transformation that would be useful and drop other useless features.

i. Derived measures

1. price/m2 - will help more obtaining more informative insights on comparisons, since we are normalizing the price by the house size, and then there is a more fair comparison.

f. Document what you have learned.

7. Solution Proposal
    a. Do Data Setup and Feature Engineering
    b. Test hypotheses that rose from Data Analysis and create an Actionable Insights report.
        i. Final deliverables:
        ii. Expected Tools:
            1. Jupyter Notebook
        iii. Tasks process:
            - H1: Houses that have a water view, are 20% more expensive on average.

              Solution: Do a boxplot to show that this hypothesis is true. In the EDA, check distributions over box plots, and confirm an insight that the 25% cheaper houses with waterview are already more expensive than 75% of all houses with no waterview.
            - H2: Houses with year built older than 1955, are 50% cheaper on average.

              Solution:
            - H3: Houses without basement are 40% bigger than with basement.

              Solution:
            - H4: The growth rate of the houses price YoY (Year over Year) is 10%.

              Solution:
            - H5: Houses with 3 bathrooms have a MoM (Month over Month) growth of 15%.

              Solution:
            - H6:

              Solution: Use seaborn regplot to see a scatter plot with a regression line, and at the side the correlation matrix. Am i able to determine that the hypothesis is true only by looking at the correlation?

    c. Houses Recommendation report
        i. Final deliverables:
            1. Build a table with house recommendations to buy or not to buy.

2. Build a table with best time periods recommendations to sell houses and for which prices.
3. Check other notebooks, but add a map with the profit density, based on the houses of each zipcode indicated to buy, and with the selling price calculated.

ii. Expected Tools:

1. Jupyter Notebook

iii. Tasks process:

1. Business Question: Which houses should be bought and for what price?

   Planning proposal 1:

   - With all the data collected, integrated and treated,
   - Create a new dataframe with median variable for each region, and the zipcode, by grouping the data by zipcode, and calculate the median for each sub-group,
   - Then merge the original dataframe with the new one, on zipcode and with a left outer join, or inner join, to return all the rows that match with the first dataframe and add the correspondent median,
   - Then rise some conditional hypothesis:
     - The houses that have a price value lower than the median and are in good conditions, can be sold for a higher price, so are good to buy.
     - The houses that have a price value lower than the median and are in bad conditions, cannot be sold for a higher price, so are not good to buy.
     - The houses that have a price value higher than the median, independently from the condition, are not good to buy and take profit.
   - Create a new variable that indicates what should be bought or not

   Data sample 1:

house id	zipcode	house price	median price	condition	Status
185074	385421	450 000	500 000	3	to buy
145879	785963	400 000	500 000	1	don't buy
145879	785963	750 000	500 000	2	don't buy

   Code sample 1:

```
df = orig_df[['zipcode','price']].groupby('zipcode').median().reset_index()
df.columns = ['zipcode', 'median_price']
merged_df = pd.merge(orig_df,df,on='zipcode',how='inner')

for i in range(len(merged_df)):
  if (merged_df[i,'price'] < merged_df[i,'median_price']) &\
     (merged_df[i,'condition'] >= 2):
    merged_df['status'] = 'to buy'
  else:
    merged_df['status'] = 'don\'t buy'
```

Planning proposal 2:

- With all the data collected, integrated and treated,
- Create a new dataframe with median variable for each region, and the zipcode, by grouping the data by zipcode, and calculate the median for each sub-group,
- Then merge the original dataframe with the new one, on zipcode and with a left outer join, or inner join, to return all the rows that match with the first dataframe and add the correspondent median,
- Create a new variable, percentage_value_below_median, i.e. (1 - house_price/percentage_value_below_median).
- Then filter the houses that have more than 0% percentage_value_below_median
- And on a descending way order the houses by percentage_value_below_median and then by condition.

Data sample 2:

house id | zipcode | house price | median price | condition | percentage_value_below_median

145879 | 785963 | 400 000 | 500 000 | 2 | 20%

185074 | 385421 | 450 000 | 500 000 | 3 | 10%

145879 | 785963 | 750 000 | 500 000 | 2 | -50%

Code sample 2:

```
df = orig_df[['zipcode','price']].groupby('zipcode').median()
                                                    .reset_index()
df.columns = ['zipcode', 'median_price']
merged_df = pd.merge(orig_df,df,on='zipcode',how='inner')

merged_df['percentage_value_below_median'] = 1 -
                    (merged_df['price']/merged_df['median_price'])
```

```
merged_df.loc[merged_df['percentage_value_below_median'] > 0,
['id', 'zipcode', 'price', 'median_price', 'condition',
'percentage_value_below_median']].sort_values(
                              by=['percentage_value_below_median',
                                  'condition'])
```

2. Business Question: Once its bought when it's the best time period to sell it and for what price?

   Planning Proposal 1:

   - Define a new variable, 'season', that is representing seasonal quarters, ['Spring', 'Summer', 'Autumn', 'Winter'], extracted from 'date' according to local data.
   - Group houses per zipcode and sub-group per season, then calculate the count of houses by id, to get a new variable oh n_houses_per_season
   - From the same aggregation calculate the median price per season.
   - Merge the two dataframes by zipcode and season.
   - Calculate median_price_per_season_per_house = median_price_per_season/n_houses_per_season. This will normalize the median_price, and the comparison between season will be fair.
   - Since the best time period granularity chosen was quarterly seasons, we must extract and add to each zipcode record from the merged dataframes the best season.
   - Then drop the season, n_houses and merge it with the original dataframe on zipcode with a inner join.
   - Then for each already listed house record per percentage_value_below_median, sort it as well per the potential seasonal selling price, with the best time period on the side.

   Data sample 1:

   house id | zipcode | house price | median price per zipcode | condition | percentage_value_below_median | best selling period | potential selling price

   145879 | 785963 | 400 000 | 500 000 | 2 | 20% | Autumn | 600 000

   185074 | 385421 | 450 000 | 500 000 | 3 | 10% | Summer | 650 000

145879 | 785963 | 750 000 | 500 000 | 2 | -50% | Summer | 550 000

Code sample 1:

Planning Proposal 2:

Ver Meigarom Proposal

Data sample 2:

Code sample 2:

Planning Proposal 3:

- Assuming that the data was already collected, integrated and treated. And the variable about 'selling_price' is available. But it depends on what attributes are available, regarding the date of purchase the acquiring price, and some records with both acquiring and sold price.
- First define a new variable, 'season', that is representing seasonal quarters, ['Spring', 'Summer', 'Autumn', 'Winter'], extracted from 'date' according to local data.
- Then group the dataframe by 'zipcode' and sub-group by 'season', and calculate the median selling price per season per zipcode.
- Then calculate the maximum median value for each zipcode sub-group, and get the seasonal quarter which has the highest median selling price.
- Then merge it to the original dataframe, on zipcode and with a left or inner join. And now we have answered which is the best time to sell, assuming that the reason the seasonal quarter had the highest median by chance, because there could be a random chance that in that time period some higher price value houses were all sold.
- And create some conditional hypothesis:
    - The houses selling price will be equal to the median + 10% if the house characteristics are better, like condition = 3.
    - The houses selling price will be equal to the median - 10% if the house characteristics are worse, like condition = 3.

Data sample 3:

house id | zipcode | house price | best_selling_quarter | median_selling_price | condition | selling_price

10330 | 857496 | 450 000 | Summer | 700 000 | 3 | 700 000 + 10%

Code sample 3:

Planning Proposal 4:

- Same as previous one but now adding the number of bedrooms and sqft_lot to the seasonal aggregation to calculate a median for houses with the same characteristics.
- And the percentage of the median selling price will be based on the YoY (Year over Year) variation. If its positive plus 10%, much higher than previous years plus 20%, etc.

3. Business question: To rise the housing selling price, the company should do a renovation. So what would be renewal changes?

   Planning Proposal 1:

   - By analyzing the impact of the different amenities and concluding for the houses that were recommended before, which amenity should be updated.
   - By seeing a visualization of price vs sqft_lot, colored by different amenities that will illustrate the boundaries of which amenity should be improved for the house with a price and a fixed size.
   - Compare by only see the price vs amenity

   Data sample 1:

   Code sample 1:

d. Create an interactive Dashboard with the statistics and maps visualization.

   i. Final deliverables:

      1. An URL for the access to the dashboard, where the business insights and results will appear.

   ii. Expected Tools:

      1. Pycharm
      2. Streamlit

   iii. Tasks process:

1. Create a main page where the dashboard shows more customization to explore the houses portfolio data, with the statistics and maps. Below each button, slider or in the section where the information will be visualized, there should be some instructions, if needed, if not it must be intuitive, to which button to use to get information.
2. In the statistics graphs try to use a list of labels to replace the numeric values for the categoric attributes, instead of transforming the features.
3. Then create a sub-page which appears on a Recommendation System section, and it show the Available Reports, and this is where the last report with the insights and business problems answered will be.
    a. Add the both reports of Houses Recommendations and Actionable Insights to the dashboard.
        - On dashboard, Actionable Insights report,only present the 5 main insights of the business hypothesis rosed.
            - Add Business Results
            - Add Actionable solutions
            - Business metrics comparison between old and newer actions.
            - Add a conclusion
                - Was the initial goal achieved?
            - Next actionable steps
4.

e. Make the interactive Dashboard web accessible.
    i. Final deliverables:
        1. An URL for the access to the dashboard.
    ii. Expected Tools:
        1. Heroku
    iii. Tasks process:
        1. Create Procfile, setup.sh, requirements.txt etc.

References

- https://github.com/lfreitas16/Insights-House-Rocket#readme
- https://github.com/feliperastelli/Projeto-Insight-House-Rocket
- Live Office Hour 09
- Live Office Hour 10
- https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices