Capstone Public Project Summary

Student Name	Hassaan Ahmed Siddiqui
Student ID	1902450
Company Name	Datapolis
Project Category	Predictive Analytics
Faculty Supervisor	Mihaly Orsos

Datapolis is a new, data-driven digital tool supporting real estate investors in investment decisions by analyzing various dimensions of cities. Furthermore, Datapolis aims to provide novel insights, comprehensive reports, interactive visualizations, and machine learning forecasts about relevant, complex urban metrics at different granularities from the city-level to the street-level.

Since the company is in its initial stage, it is working towards building a database from scratch. The client requires data for rental apartments from multiple cities that could then be cleaned and used for further analysis and predictive modelling. Therefore, I am working towards scraping data from websites where I can get data for rental apartments/houses. In addition, I am also trying to run some statistical analysis and predictive modelling for each city and all cities combined to understand the difference in importance of variables from city to city and on an overall level. This project aims to

i) Collect aligning to the websites' privacy policies and regulations, publicly available data about rental apartments prices from long term rental websites for different cities.
ii) Organize the collected data for each city into a database.
iii) Do data cleaning procedures and feature engineering to make sure that data is ready for further analysis.
iv) Provide exploratory data analysis aggregated on the level of each city and all cities respectively.
v) Design predictive models on the rental prices and evaluate the performance.

Data Acquisition:

I acquired the data using web scraping technique in which I used a programming language "*Python*" to retrieve the data for rental apartments and convert the data into a computer readable format which can be further used for analysis. During the process, I managed to select a website which I could use to scrape data from.

Data Cleaning:

Once I successfully acquired the data. The next step was to clean the data to make it ready for Exploratory Data Analysis and Modelling. The cleaning process includes:

1. Removing duplicates.
2. Removing weird characters and values from Amenities column.
3. NAs were removed from selected columns.
4. Cities with multiple names was fixed.

Feature Engineering:

Amenities were important therefore I created a column for each individual amenity and converted them into binary values.

High ordered variables were also created. (square, cubes and logs) were calculated from numeric variables as we have seen in the last section. The main purpose of this engineering is to make sure that we use the best distribution in the modelling and predictions.

Quartiles were created out of prices into 4 classifications Q1, Q2, Q3 and Q4. Prices of the apartments were divided into four quartiles based on mathematical distribution principal.

Distance from the city center was calculated using longitude and latitude of apartments which was provided in the data. The longitude and latitude for city centers of each city was taken from the google maps. Finally, Using the function of distGeo I calculated the distance of each listing from the city center.

Exploratory Data Analysis:

When data was cleaned, I did some exploratory data analysis to see the relationship of the predicted variable with the predictor variables. There were two continuous variables which I found to have an interesting relationship with price. Therefore, I plotted dot-plot of price with distance from city center and price with apartment size. On the same plot I ran Loess regression, the result of which was smooth line showing the overall movement of price

I also looked at the distribution of each city with respect to price, Distance from city center and apartment size. Boxplots for each variable were plotted and distribution was evaluated, and insights were drawn out.

Predictive Modelling and Analytics:

Before delving in the model building and predictions I did converted the predicted variable which is price in this case into different classes which were based on quartile methodology as shown in the feature engineering section. For the purpose on modelling I made two types of classifications:

1. Binary which consisted of only quartile 1 and 4.
2. Multiclass classification which consisted of all the classes.

Model Selection:

I selected Random forest as an algorithm to be used for predictive analytics. It is because I was using classifications and Random Forest is use for classifications. Apart from its essence it is less influenced by outliers than other algorithms. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, it won't allow overfitting trees in the model.

Performance and Results:

For Binary classification Random forest has performed exceptionally well with an average accuracy score of more than 90%. The accuracy measure of other performance metrics has also been great.

Distance from the city center has been one of the most important variables for most of the cities in term of its relative importance to the decision making of the algorithm.

Amongst all the individual amenities, washing machine was the most consistent amenity to be present in the top 10 important variables for all the cities evaluated individually.

As compared to Binary, the performance of multi class was low. On average, the accuracy of the model was hovering between 60% to 70%. The RMSE values were consistent between 0.53 to 0.55. R squared values were ranging from 0.7 to 0.75.

Distance from the city center and apartment size were the top two variables both for all cities combined and individual cities. Individual amenities were different for the individual city.

Recommendations:

1. The scope of the project should not be limited to only long-term rentals. It should be expanded to short term rentals and commercial sites as well.
2. All the ongoing projects should be linked together to see if a meaningful story is can be drawn.
3. This project can be used to see which cities are similar in nature with respect to their price distribution and variable importance. This will further help the client create a group of cities showcasing similarities. Based on this observation further recommendations can be drawn.
4. The prices used reflect abnormalities because of pandemic. It is advised that past data should also be used and check the difference in the price patterns.

Learning Outcomes:

1. Developing and improving skills on collecting large-scale urban data through web scraping.
2. Preprocessing, cleaning, and organizing data.
3. Building statistical and machine learning models in R.
4. Gained deeper understanding and working experience on real-estate development.
5. Practice communication and teamworking skills.
6. First-hand experience of the product development process of an early-stage start-up.