Personalized Recommendation System

Ashera Dyussenova
BS20-DS-01
Innopolis University
Innopolis, Russian Federation
a.dyusenova@innopolis.university

***Abstract*—CRISP-DM (Cross-Industry Standard Process for Data Mining) outlines a six-phase data mining life cycle, which ranges from identifying an objectives from a business perspective till putting a technical solution into reality. CRISP-DM starts with Business Understanding stage, where define project goals and business requirements. Knowledge is transformed into a problem description for data mining and a rough plan for accomplishing the project objectives. On the next step, we examining the data that is available for mining. It involves exploration of data using different tools such as tables and graphics to determine the quality of the data. Data preparation involves refining and modifying raw data before it undergoes processing and analysis. This crucial step involves tasks such as reformatting data, rectifying errors, and merging data sets to enhance the quality of data. Modeling is the fourth stage, which involves developing models using selected variables and data prepared in previous steps. Different modeling techniques can be used, and the model is evaluated for its usefulness in solving the business problem. And on the last stage of Evaluation we are comparing results with initial goals.**

In this article, we analyze Airbnb dataset to establish personalized recommendations system for marketplace companies.

Keywords—CRISP-DM, Business Understanding, Data Mining, Business Objectives, Data Understanding, Data Preparation, Modeling, Evaluation.

I. INTRODUCTION

As the Internet has grown in popularity, more and more businesses are integrating their products to the web. Airbnb is one of the online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales. The business has advanced significantly since 2007, when its co-founders first had the notion to allow paying visitors to spend the night on an air mattress in their living room. The most recent data from Airbnb shows that company has more over six million listings worldwide, covering more than 100,000 cities and towns.[1]

In addition, the HORECA (hotel, restaurant, and catering) sector has been significantly impacted by Airbnb. According to some industry experts, Airbnb's explosive rise has upset the conventional HORECA model and, in some markets, has resulted in a drop in income and occupancy rates for conventional hotels.[2]

Due to the extensive longevity, and other arguments given upper, Airbnb has been selected as the foundational basis for our model to create universal personalized recommendation system in this business field.

II. BUSINESS UNDERSTANDING

The Business Understanding stage is where we develop specific understanding of the business processes. And where we try to concretely define the problem we will working on. This stage is primarily about strategically aligning business processes with execution of the project. Correctly understanding the business functions could save time, so not jumping into solution step which builds on incorrect assumptions.

Let's break down Business Understanding phase into several parts: Business Objectives, Assess Situation, Determine Data Mining Goals and Produce Project Plan.

A. Business Objectives

Businesses are facing more competition from brand-new websites as more HORECAs (hotel, restaurant, and catering) start selling their products online. A business must discover strategies to stay profitable while it is up against competition. One of the proposed solutions is development of an individual approach to each client in order to maximize the fulfillment of the needs of the company's clients.

Thus, the main objective is:

- Improve sales by making individual recommendations for specific client.

(A) *Business Success Criteria*

To deemed study successful it should meet several aspects such as:

- Increasing of sales by 8-10% in 1 year
- Increase in time spent and pages on the site by 5-10% in 1 year
- Adhere to the allocated budget and timeline.

B. Assessment of the Situation

To ensure that the data mining project is executed effectively and efficiently, it is essential to conduct a thorough assessment of the available resources, constraints, assumptions, and other relevant factors. This information is crucial in determining the adequacy of the company's resources for the data mining process and other subsequent phases. In this regard, a detailed inventory of resources needs to be compiled, including personnel, data sources, computing and software tools.

The personnel required for the project will comprise two business experts, three data experts, two technical support professionals, and three data mining personnel. The data sources to be used for the project will be drawn from the database, which is based on the information of customers who have already registered on the site. Computing for the project will be facilitated by the hardware computing owned by company members. The software tools required for the project will include Microsoft Office for business experts, SQL and Google Colab for data experts and mining.

Additionally, the project's requirements need to be identified to determine whether the company has adequate resources to undertake the data mining process effectively. The project will require the hiring of additional personnel, including a cybersecurity expert, lawyer, and accountant. Moreover, it will be necessary to acquire office space equipped with the appropriate equipment for the project's execution. Technical aspects such as data legality, scheduling with deadlines, and quality of work requirements will also need consideration.

Based on the client experience preferences in certain areas, it is feasible to offer similar results to aid in consumer satisfaction. However, the project's constraints include the possibility of unavailable data and the need to verify the legality of data, which may affect the project's outcomes. Therefore, it is important to identify the risks and contingencies for the project, including the main risks of time spent on research and expenses, with appropriate mitigation strategies in place. In summary, a comprehensive assessment of the available resources, constraints, assumptions, and risks is necessary for the successful execution of the data mining project.

C. Data Mining Goals

To make aims more clear it is important to convert company's business objectives into data mining terms. Based on the business goal, I want to highlight two main **Mining Goal:**

To build a model relating "similar" things, use historical data from earlier purchases. Link to other things in the relevant group when users read an item's description.

D. Project Plan

After elicitation of objectives and goals, it is time formulate concrete stages and deadlines. The project plan serves as the primary document for your entire data mining endeavor. It provides information on the objectives and timetable for all phases of data mining to everyone involved in the project.

Plan overview			
Phase	***Time***	***Resources***	***Risks***
Business Understanding	5 week	All analysts	Economic change
Data Understanding	5 week	All analysts	Data problems, technology problems
Data Preparation	8 week	Data mining consultant, some database analyst time	Data problems, technology problems
Modeling	6 week	Data mining consultant, some database analyst time	Technology problems, inability to find adequate model
Evaluation	4 week	All analysts	Economic change, inability to implement results
Deployment	4 week	Data mining consultant, some database analyst time	Economic change, inability to implement results

III. Cost and Benefit

To calculate Economic Benefit we could use Benefit/ costs analysis. Benefit/costs analysis is one type of economic valuation – an analysis that assesses the relative value of a project in monetized estimates. As the name implies, benefit/ cost analysis determines the value of a project by dividing the incremental monetized benefits related to a project by the incremental costs of that project. The result is called the **Benefit/Cost Ratio** and is often the primary output of the analysis process.[3]

	$
Benefits	500 000 $
Costs	100 000 $
B/C Ratio (Benefits/Costs)	5.0
Net Benefit (Benefits-Costs)	400 000 $

According to statistics, the average yearly earnings of an Airbnb is over $5 000 000. If project could work, then the 10% of average is $500 000. Approximate expenses $100 000.

IV. Data Understanding

In the second phase of the Cross-Industry Standard Process for Data Mining (CRISP-DM), after appreciation of business goals and data mining plan, it is time ti start working with data itself. Our first step is gather the data with defined selection criteria, verify data availability and outline data requirements. Using basic statistical techniques examine the data more closely, estimate minor and major quality issues. Lets split Data Understanding phase to 4 main parts: Initial Data Collection, Data Description, Data Exploration, Data Quality. [4]

A. Initial Data Collection

Proper data is essential to the success of the entire project. In this paper we will use already collected data from Airbnb listings, so at the current moment there is no need to purchase external databases or spend money conducting surveys.

As initial data we have 4 tables: Listings data dictionary.csv, Listings.csv, Reviews data dictionary.csv, Reviews.csv. The database is in appropriate format and it is compatible with data-mining platform.

To load data to environment, it should be encoded using "latin_1" encoding method due to presents of other latin languages except English.

B. Data Description

In the dataset there is four data-frames. Two main and two description of attributes of main datasets. Two main data-frames overlap only by one column listing id. First main dataset has 4 attributes, we will need only two of it: listing id and responding reviewer id. Second main dataset has 33 attributes and describes listing itself. The most important columns are location including district, city, in the form of string and longitude, altitude as floating numbers. Another important columns are different reviewer scoring in form of floating numbers in range of 0 and 10. And column in which we most interested amenities in text format. Further we will reform amenities column and will focus on it.

- **Listings_data_dictionary.csv -** present columns of listing.csv and short review of content of columns.
- **Listings_data_dictionary.csv -** main data-frame which contain 33 columns with detailed description of listing. Columns that we are going to focus are:

listing_id, price and amenities. Listing_id will present as key identifier for certain listing. Price and amenities are useful features.

- **Reviews_data_dictionary.csv -** present columns of reviews.csv and short review of content of columns.

- **Reviews.csv -** secondary data-frame with 4 columns which present review information. We are interested in reviewer_id, listing_id and date columns. Reviewer_id have role of key for certain person, connected to listing_id. And Date column will be used in time series analysis.

C. *Data Exploration*

Our main focus is on amenities column. It is datatype is string. String consist of keyword, that has listing included. For example: Iron, Wi-Fi, Microwave, etc. Our hypothesis based on preference of client, so in the next phase we will extract items and present them as attributes, so it is easier to paint a picture of preference of client.

In Figure 1, the "Amenities" column in the dataset is a categorical variable with 245,003 distinct values, accounting for 87.6% of the total observations. There are no missing values in this column.

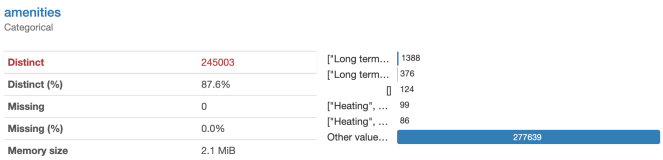


Figure 1.

Another important features are listings id and reviewer id. The columns represented in form of integer number without repetition. Matching client with listings that he had will open the key preference.

In the figure 2, the "reviewer_id" column in the dataset is a real number variable with a mean of 98,081,330.42. It has 4,450,005 distinct values, accounting for 82.8% of the total observations. There are no missing values, infinite values, or negative values in this column. The minimum value in the column is 1, and the maximum value is 390,338,478. There are no zero values in this column.

Figure 2.

In the Figure 3, The "listing_id" column in the dataset is a real number variable with a mean of 16,029,886.47. It has 193,556 distinct values, accounting for 3.6% of the total observations. There are no missing values, infinite values, or negative values in this column. The minimum value in the column is 2577, and the maximum value is 48,263,869. There are no zero values in this column.

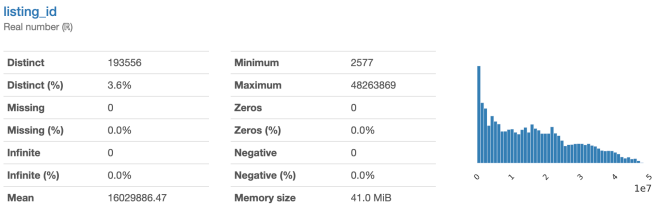


Figure 3.

D. *Data Verification*

Data represented earlier does not contain any duplicates. No listing id has null value, so all id are present. Column Amenities has at least one item present, empty rows considered as 0 presents and might be excluded from dataset. Overall, dataset contains all needed information.

V. DATA PREPARATION

The third stage of CRISP-DM, Data Preparation will be the subject of our focus. Once the data has been identified and comprehended, the process of data preparation becomes necessary, which includes activities such as data cleansing, data integration, data encoding, feature engineering, as well as variable and feature selection. During this stage, we will transform the data into a format that can be effectively utilized in the subsequent modeling phase. Data Preparation consists of 5 main parts: Data selection, Data cleaning, Data constructing, Data integration and Data formation.

A. *Data selection*

The selection of data for analytical purposes involves making a decision on the basis of several factors. First, the data should be relevant to the goals of the data mining process. Second, it should be of good quality, meaning that it's accurate and reliable. Finally, there may be technical constraints to take into account, such as limits on the amount of data that can be processed or the types of data that can be used.

Our main goal is improve sales by making individual recommendations for specific client. Based on this statement we could get rid of columns about location(district, city, latitude, longitude), because traveling of certain person usually changing. Also information about host itself useless, clients are not interested in such details. According to Data Understanding stage, rating columns not reliable and contains missing values.

Column Amenities is the most suitable for building the individual recommendation model. Another important features are listings id and reviewer id, which would be used for making mapping between certain listing with amenities preferred and person who choose it.

B. *Data Cleaning*

After selecting appropriate features it is step to improve the quality of the data so that it is suitable for the analysis methods we have chosen. Cleaning data using default values where appropriate, or using more advanced techniques like modeling to estimate missing data.

We start with examining amenities column for null or missing value. Because, this column is represented as object text, there is no null value, but there exist just empty strings. To clean it, we are dropping all rows with text length less

than 30 characters, because it is too short to present any useful information.

Second, listing id do not have null or missing values. It is not unique, because different reviewers could access same listing.

Finally, reviewer id column is in the same state as listing id and matches it in length.

C. Data Constructing

This task involves working on data to make it useful for analysis or processing. It involve creating new data from existing data, adding new records, or changing the values of existing data to make it more suitable for the intended purpose.

At this juncture, our focus is on processing the amenities column. This column contains a list of items present in a specific apartment, such as wifi, TV, parking, among others, in text format. Our approach involves extracting each item individually and subsequently tallying the frequency of their occurrence in the dataset.

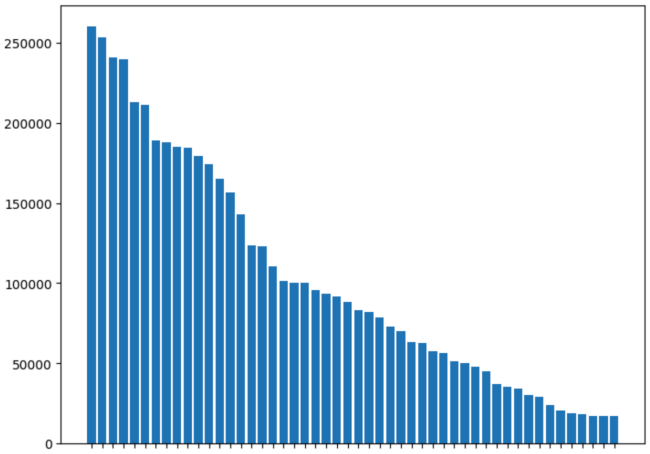


Figure 4.

Figure 4 x-axis demonstrates most frequently appeared top 50 items in dataset and y-axis the number of appearance in dataset. Based on figure 1 there are most important items occur 10 000 times, and after entire item count heavily falls. Additionally, it is better to add few more items to have more data.

After detecting base items, we convert it into features which would be main data for our modeling. New features will contain values of 0 and 1. If this feature is present in the certain listing, then it is marked as 1 and 0 otherwise.

D. Data Integration

These method refer to ways of taking information from different sources, such as tables or databases, and combining them to create new complete ready dataset.

Now we have dataset1 with valueless columns and new 30 useful features connected with listing id. Also we have dataset2 with listing ids and reviewers id connected together, we should drop columns and merge databases to one.

Finally, we got ready one final dataset with reviewer ids, listing id that they have accessed and features that present current listing.

E. Data Formation

Format data are changes made to the way data looks, such as how it is structured or organized, but these changes do not change the actual meaning of the data. These changes are often needed so that the modeling tool can properly analyze the data.

Model will learn based on reviewer's preferred features and fill search from dataset con-similar listings and recommend them. Data formulated in previous stage is appropriate and do not needed in formation.

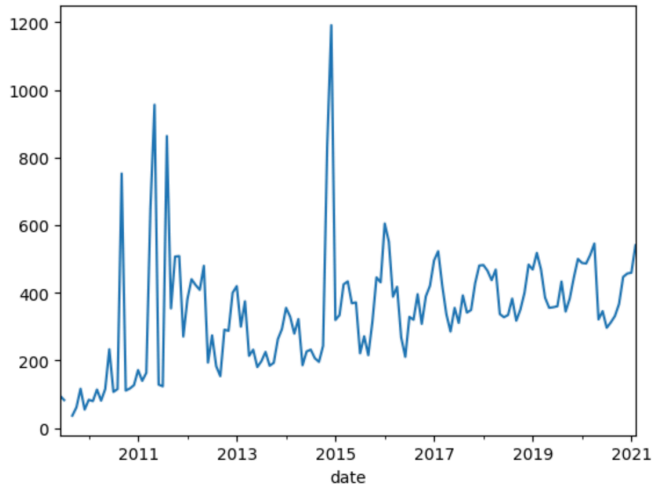
VI. MODELING

In this stage, the modeling tool was utilized with the datasets that were created and prepared in the previous stages. The main goal is to find the model that can effectively explain the existing data and predict the target variable for new data. Modeling consists of 4 main stages: Select Modeling Technique, Generate Test Design, Build Model and Assess Model.

Since the aim of the study was to predict listings, an unsupervised machine learning algorithm model was needed. Clustering is algorithm that groups data points into distinct clusters. It is a method of data analysis that can be used to identify patterns and relationships in data.

A. Time Series Analysis

Firstly, for the purposes of time series analysis we resample the dataset with 500000 random rows. After convert the date to a format that supports TSA(time series analysis). Among all the features there is only one non-binary that can be meaningfully represented in a time series analysis - **price.** After drawing time series it was hard to understand, so we decided to find the mean over a month.



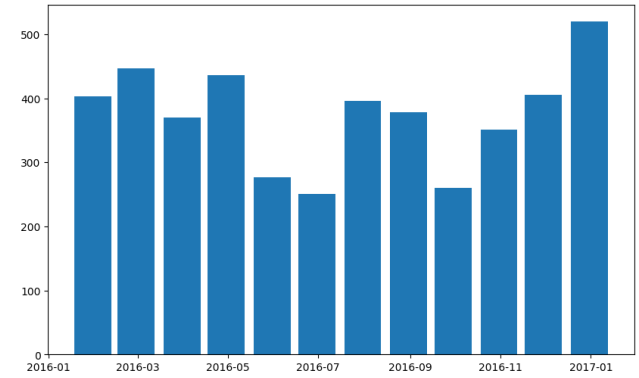
With time series presented upper we can already make some inferences:

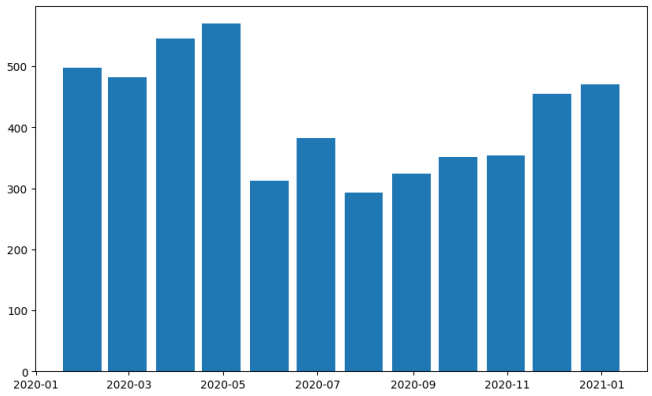
1. **Trend**: we can see that prices are increasing linearly over time, excluding jumps in prices in 2011-2012 and some of 2015. This can be explained by inflation, and ideally it would have been accounted for in the dataset - that is, prices would have been resampled to make sure that features in more recent years are not more valuable to clusterization algorithms than those of the older ones.

2. **Seasonality**: on the graph we can also notice the seasonality of price increases - prices jump up closer to

the beginning of the year - that could be explained by Christmas and New Year celebrations. The original dataset also did not account for that, so seasonality might hamper the process of clusterization, pulling apart apartments that were often presented at New Year and those that were more often provided during summer.

Seasonality becomes more evident if we look at mean prices in the scope of a single year:





We can see that seasonality holds even during a major crisis, such as COVID-19, which initiated a wave of lockdowns and significantly harmed the business model of Airbnb, confirming its great importance for the data.

Conclusion: both trend and seasonality are very important for the price feature, and have to be taken into account in the future for more effective clusterization.

B. *Select Modeling Technique*

The initial step in the modeling process involves the selection of an appropriate modeling technique. It is imperative to choose the most suitable technique for the given task. The commonly used models for clustering algorithm type of problem include k-means clustering, hierarchical clustering, and density-based clustering, Expectation-Maximization (EM) Clustering, Gaussian Mixture Models (GMM), Affinity Propagation and etc.

We are going to focus on three most popular clustering algorithms, which are simple in implementation and effective in usage:

- **K-means clustering** is used to group data into clusters. It works by assigning each data point to the closest cluster based on its distance from the cluster's center. The algorithm then iterates until the clusters are optimized and no further changes can be made. The number of clusters is predetermined by the user and is denoted by the letter k.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** is used to identify clusters of data points within a dataset. It is a density-based clustering algorithm that uses a two-step process: first, it identifies core points (points that have more than a specified number of neighbors) and then it expands clusters from the core points.
- **Agglomerative clustering** starts with each data point as its own cluster and iteratively merges the two closest clusters until a stopping criterion is met. The distance between clusters can be measured using various metrics, such as Euclidean distance or cosine similarity. The algorithm produces a dendrogram, which is a tree-like diagram that shows the hierarchical relationships between the clusters.

C. *Generate Test Design*

In order to conduct a comprehensive analysis of the performance of a predictive model in the context of a real-world application, it is often necessary to generate test data that accurately reflects the characteristics of the underlying population. In this regard, the current study has utilized a set of pre-defined features, including 'listing_id', 'reviewer_id', 'price', and a range of amenities such as 'Wifi', 'Essentials', 'Kitchen', 'TV', and 'Air conditioning', among others. The resulting test data, comprised of a random sample of 10% of the full dataset, is intended to facilitate a rigorous evaluation of the performance of the predictive model under consideration.

D. *Build Model*

K-Means algorithm

Model performs clustering on a dataset of listings using K-Means algorithm to group similar listings together based on their features. It then recommends a list of similar listings to a given reviewer based on the cluster of the listing they have already reviewed.

The dataset is preprocessed using StandardScaler to standardize the numerical features of the listings. Then, K-Means clustering algorithm is applied to the dataset with the number of clusters ranging from 1 to 20. The Within-Cluster-Sum-of-Squares (WCSS) is calculated for each value of k to determine the optimal number of clusters. The elbow method is used to visualize the WCSS values against the number of clusters, and the number of clusters is chosen based on the point where the change in WCSS value becomes negligible.

In figure 5, we could see result of Elbow Method. Based on this result number of clusters was chosen 12.

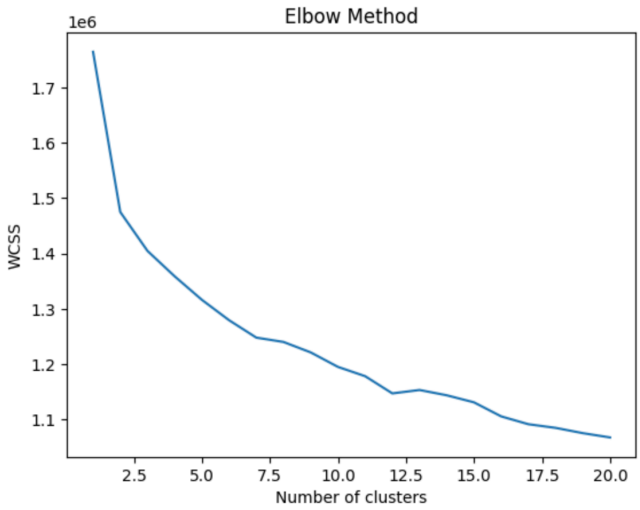


Figure 5.

After clustering, a function named `recommend_listings` is defined which takes a reviewer ID and the dataset as input, and returns a list of recommended listings for that reviewer. The function first identifies the cluster of the listing reviewed by the given reviewer and filters the listings in the same cluster. It then excludes the listing already reviewed by the given reviewer and returns a random sample of similar listings from the filtered list based on the value of n_recommendations parameter.

AGGLOMERATIVE CLUSTERING

Hierarchical clustering algorithm. The features of the listings are standardized using StandardScaler to ensure that all features have the same scale.

Then, hierarchical clustering algorithm is applied to the standardized features with the number of clusters set to 5. The 'euclidean' distance metric and 'ward' linkage criterion are used to measure the dissimilarity between the listings and merge the clusters. The predicted cluster labels are added to the dataset.

The dataset is then sorted by date to visualize the changes in price over time. A scatter plot is created to display the clusters of listings over time, where each cluster is represented by a different color. (Figure 6)

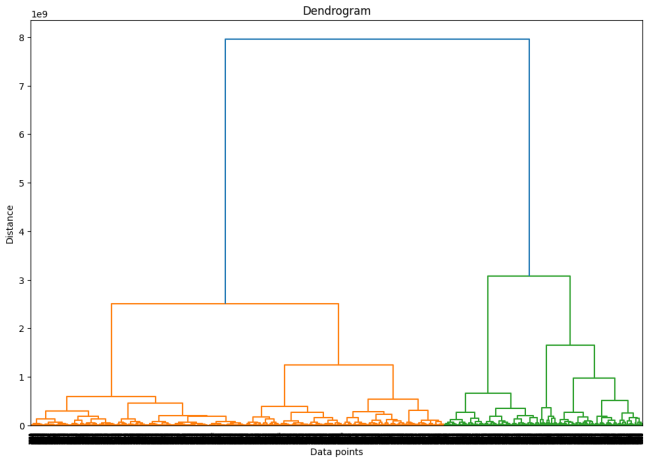


Figure 6.

DBSCAN

Model using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, features of the listings are standardized using StandardScaler to ensure that all features have the same scale.

Then, DBSCAN algorithm is applied to the standardized features with the hyper-parameters eps and min_samples set to 0.5 and 5, respectively. The 'euclidean' distance metric is used to measure the dissimilarity between the listings. The predicted cluster labels are added to the dataset.

Unfortunately, result of DBSCAN are very poor ad represented in Figure 7. Almost all clusters are -1.

Cooking basics	Elevator	Bed linens	Microwave	Fire extinguisher	Dryer	Stove	Coffee maker	Oven	First aid kit	Carbon monoxide alarm	Free parking on premises	Private entrance	cluster
1	1	0	1	1	0	1	1	1	1	1	0	0	-1
1	0	1	1	0	0	0	0	1	0	0	0	0	-1
1	0	1	1	1	0	1	1	0	0	1	0	1	-1
0	0	0	0	0	0	0	0	0	0	0	1	0	-1
1	1	1	1	1	0	1	1	1	1	1	0	0	-1

Figure 7.

E. Assess Model

Among the three clustering algorithms applied on the dataset of listings, agglomerative clustering performed the best. It successfully grouped similar listings together based on their features and provided insights on the changes in price over time. The elbow method was used to determine the optimal number of clusters, and the 'euclidean' distance metric and 'ward' linkage criterion were used to measure the dissimilarity between the listings and merge the clusters. The resulting clusters were visually represented in a scatter plot with each cluster represented by a different color. On the other hand, K-Means algorithm also performed well and provided a recommended list of similar listings to a given reviewer based on the cluster of the listing they have already reviewed. However, DBSCAN algorithm did not perform well and resulted in almost all clusters being labeled as noise.

V. EVALUATE RESULTS

A. Evaluate Results

The assessment results were evaluated in terms of business success criteria. The data mining results were interpreted and checked against the given knowledge base to ensure the discovered information was useful and novel. The evaluation and assessment were compared to create a ranking of results based on business success criteria. The impacts of the results on the initial application goal were also checked. Overall, the agglomerative clustering algorithm performed the best, grouping similar listings together and providing insights on changes in price over time. The K-Means algorithm also performed well by recommending similar listings to a reviewer. However, the DBSCAN algorithm did not perform well and resulted in almost all clusters being labeled as noise. The project's success in meeting initial business objectives was not mentioned.

B. Review Process

Two potential limitations were identified in the current study.

The first limitation pertains to the lack of data, which could potentially affect the accuracy of the classification results. As such, additional datasets may be necessary to further enhance the reliability of the findings.

The second limitation concerns the potential benefit of supplementing the existing data with other relevant features to achieve more accurate classification. In doing so, the predictive power of the model could potentially be increased, resulting in more reliable and valid predictions.

These limitations highlight the importance of ensuring that sufficient and relevant data is available, as well as considering the potential benefits of incorporating additional features to improve the performance of predictive models. Future research could explore these limitations further and devise appropriate solutions to address them.

C. Determine Next Steps

We need to make more improvements before we can use the data mining results for business purposes. Our next steps will be to fix the problems we found earlier and run the clusterization process again.

At the onset of the data mining process, it is imperative to perform an initial assessment of the available dataset to determine its adequacy for the intended purposes. In the event that the dataset is deemed insufficient or incomplete, it may be necessary to procure additional data to supplement

the existing dataset. Additionally, it is essential to review the features of the dataset to ensure that the most relevant and appropriate variables are included in the data mining process. Thus, the initial steps in this process entail an evaluation of the dataset's suitability and completeness, as well as a comprehensive review of the dataset features to ascertain their relevance to the intended data mining objectives.

REFERENCES

1. "Company Overview of Airbnb, Inc". Bloomberg L.P. Archived from the original on January 8, 2018.
2. Chen, Yong (May 13, 2021). Economics of Tourism and Hospitality: A Micro Approach. Routledge. ISBN 978-1-000-37238-0.
3. Benefit/Cost (B/C) Analysis, United States Department of Transportation, June 18, 2020.
4. "Data Mining for Dummies". Meta S.Brown, September 29, 2014.
5. "Airbnb Listings & Reviews", Mysar Ahmad Bhat, kaggle, 2021. https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews
6. IBM, IBM SPSS Modeler CRISP-DM Guide documentation. https://www.ibm.com/docs/en/spss-modeler/saas?topic=guide-introduction-crisp-dm
7. ISCTE IUL, Department of Information Science and Technology, A Text-Mining based model to detect unethical biases in online reviews: A Case-Study of amazon.com
8. "The CRISP-DM Discussion Paper", Pete Chapman, Randy Kerber, Julian Clinton, March, 1999.