Data Mining and Machine Learning I: Paris Housing Classification, US Airline Tweet Sentiment Categorization, and German Rent Projection

Viplav Vijay Gadewar
School of Computing
National College of Ireland
Dublin, Ireland
x21164274@student.ncirl.ie

***Abstract*—Only 2.2 million people reside within the "city walls" of the Paris area, which has a population of around 12 million people. Paris is a small city, measuring little over 10 kilometers north to south and roughly 12 kilometers east to west, and space is limited. As a result, renting an apartment in Paris is relatively expensive, especially when compared to other cities. According to World Air Transport Statistics, passenger demand increased significantly in 2019. In 2018, revenue passenger kilometers grew 7.4% over 2017. In 2018, China added one of most passenger trips, and the US domestic market retained the nation's largest passenger market. To enhance income and profit, most airlines devised tempting packages and high-quality services for their customers.The German home market exploded in 2020 and 2021, thanks to immigration, pandemic stimulus, European market expansion, low interest rates, and pandemic housing adjustments. And the outlook is for considerably more expansion. The German property market has been heating up in recent months, with more investment flowing in.**

***Index Terms*—Random Forest Classifier, Decision Tree classifier, KNN (K-nearest neighbour) Classification, Random Forest Regression, Linear Regression, Naive Bayes Theorem.**

I. INTRODUCTION

Paris is a compact city with limited area, stretching little over 10 kilometers north to south and around 12 kilometers east to west. As a result, compared to other cities, renting an apartment in Paris is rather expensive. I feel that basic and elegant housing should be separated in Paris.

The Domestic market in the United States maintained the world's largest passenger market in 2018. To maximize revenue and profit, most airlines designed appealing packages and elevated services for its customers. The bulk of the decisions are founded on the foundations of big data analytic. Data analysts, especially those wishing to work for airline companies, must be ready to do emotion analysis on data collected from a variety of social media sites.

Because of immigration, pandemic stimulus, European market expansion, cheap interest rates, and pandemic housing modifications, the German housing market surged in 2020 and 2021. And there's a lot more expansion on the way. In recent months, the German property market has gotten hotter, with more money streaming in. The facts and numbers below show that there is a notable trend. The scarcity of rental apartments in Germany adds to the allure of buying real estate in Germany. Construction costs and a lack of sufficient land are major roadblocks. As the European and German economies recover, we should expect rising rent prices in the main cities of Berlin, Munich, Frankfurt, and Stuttgart. The German market is attractive to all real estate investors.

Some real estate businesses feel that Germany is the best market in Europe to invest in. Four German towns were awarded the greatest performers in Europe by PWC. In this piece, we'll look at the benefits and drawbacks, as well as discover more about what drives the German rental market. The overall housing stock in Germany is around 40 million dwellings, with more than half of them being rented. This makes it appear to be a good location for investors in rental homes. In Germany, rental housing accounts for 55 percent of all residences (22 million houses and flats). Multifamily structures account for 90 percent of all rents.

A. RESEARCH QUESTIONS

1. How can we classify category (Basic and Luxury) of housing in Paris?

2. How precise can we make a system that predicts an apartment's rental price?

3. How well can we anticipate customer satisfaction with airline service?

II. Related Work

We can quickly handle real-world problems with the aid of machine learning methods, resulting in efficient and fruitful solutions. However, there are downsides to having advantages. On the basis of precise findings, we may select the appropriate model. To do this, we must run the data through every algorithm available.The majority of the concerns are related to training and testing data.Because the quantity of data is so huge, it might take a long time to repair. We enter all of the data into the model and calculate the outcome for evaluation.

Various attempts have been done in the field of machine learning to categorize income into distributed categorization on the data set we chose. The names of some of the researchers who have previously worked on this problem statement are listed below and cited at the conclusion of the report.

- To anticipate the price of a single home, Sifei Lu et al. [1] suggested a hybrid model combining Lasso and Gradient regression.
- To anticipate the value of land and houses, Muhammad Fahmi Mukhlishin et al. [2] employs a variety of approaches. This study analyzes Fuzzy logic, Artificial Neural Networks, and K-Nearest Neighbor to discover the best approach for determining the pricing of a seller.
- Atharva Choogle et al.[3] Data mining techniques have been used to introduce house price predictions. It gives a description of prediction markets as well as existing markets, which may be used to create helpful forecasts and get a better knowledge of the market. As a result, it is vital to forecast the most cost-effective price for real estate consumers based on their budgets and goals.
- In Malang, East Java, Ruth Erna Febrita et al.[4] offer a data-driven fuzzy rule extraction for housing price prediction. The K Means clustering approach is utilized in this way to extract starting values in order to create fluid membership functions and inference rules for multiple residential groups.
- On the Twitter platform, Goyal and Anuranjana [5] (2019) suggested a methodology based on multiple matrices to assess and detect positive and negative responses to Demonetization.
- Singh et al. [6] (2019) investigated the overall impact of GST implementation and Indian citizens' perceptions of GST. The author of this article separated the entire period into three parts: I before GST, ii) during GST, and iii) after GST. During these three decades, we can see that 62.96 percent of states and union territories supported or were neutral on the GST. During the GST, 55.55 percent of states and union territories supported or were neutral, but after the GST, 85.18 percent of states and union territories supported or were neutral.
- Different machine learning techniques have different accuracies for different data sets, according to Fouad, Gharib, and Mashat [7] (2018). Ensemble classifier has a 93.94 accuracy for Sanders data sets, whereas SVM has a 92.71 accuracy. The Stanford 1-k datasheet ensemble classifier has an accuracy of 78.70 percent, whereas SVM had an accuracy of 78.10. The Nave Bayes classifier has an accuracy of 85.09 for HCR datasets, whereas the ensemble classifier has an accuracy of 84.75.
- Hasan and Moin's study [8] (2018) uses a mixed machine learning strategy for sentiment analysis. This research compares sentiment analysis approaches in the exploration of political viewpoints using supervised machine-learning techniques such as Naive Bayes and Support Vector Classifier (SVC).
- Aziz, Zaidouni, and Bellafkih[9](2018) investigated the design and performance of two different executions.
- The log cost and the building in terms of the number of rooms, parking spaces, baths, fireplaces, storey, and overall living area have a close relationship based on square feet [10].
- This -analysis focuses on the use of machine learning algorithms in the real estate sector. The study's goal is to predict home prices based on homes in Malang utilizing regression and particle swarm. Particle swarm optimization is used to assess the effective variables, and regression analysis is used to get the optimal coefficient of prediction. The findings of this study have been validated for use in combined prediction and particle swarm [11].
- In this article, the support vector regression model is used to predict China's housing prices from 1993

to 2002. The hyperparameters are modified using the vector support regressor model. Only 4% of the time, the genetic algorithm yielded error ratings that were lower than average [12].
- This research presents a rental forecast model which includes the Detailed Floor Image in order to see if the image has an impact on the rental forecast accuracy. The proposed model assesses image characteristics using principal component analysis, blends them with attribute vectors, and employs either linear regression or a support vector regression model. The prediction accuracy, or mean squared errors between projected and actual rentals, is computed using 1,089,090 properties from the lifull home dataset, a Japan estate service provider [13].
- According to an assessment of the relevant literature on residential units [14], age, amenities, utility, physical qualities, and position all had an impact on rental pricing.
- This paper is an example of an apartment renting research paper. As earlier research on apartment renting illustrate the relevance of location [15], the trend is moving further toward modeling the spatial features of rentals.
- This study uses four regression models and a create multiple that blends K-Nearest Neighbours and Random Forest Method to predict the apartment's worth. The ensemble approach generates prices with a minimum inaccuracy of 0.0985 and remains constant when utilizing principal component analysis. The four regression approaches used were regression analysis, support vector, random forests , and k-nearest neighbour [16].
- This study examines the performance of five distinct hedonic model parameters for transient and spatial designs in the real estate rental market, with varied degrees of complexity. The error terms are not larger than 18 percent and 13 percent, respectively, in terms of root mean square error of approximation and mean absolute error [17].

III. IV. Data Mining Methodology

The KDD Process is a popular data analysis life cycle that tries to discover essential trends and patterns while removing 'noise' (unwanted outliers). The outcomes are listed in the table below. To examine them, we employed three different data sets and six different models.

A. KDD - Methodology

I. "Paris Housing Classification"

a. Data Selection:
This is a set of statistics derived using fictitious data on house prices in a city - Paris. This dataset is used for educational purposes, as well as practice and knowledge acquisition.

It has 18 columns and 10,000 rows represented in Table I.

TABLE I
PARIS HOUSING CLASSIFICATION

Column Name	Description
squareMeters	Area in square meter
numberOfRooms	No of rooms
hasYard	has yard?
hasPool	has pool?
floors	Number of Floors
cityCode	zipcode
cityPartRange	city range
numPrevOwners	number of pre owner
made	made in year
isNewBuilt	wheater new build
hasStormProtector	has storm protector
basement	has basement
attic	has attic
garage	has garage
hasStorageRoom	has storage room
hasGuestRoom	has guest room
price	price
category	category of house
Target Variable	category

b. Data Pre-processing:
The data set is described in the following figure 1. It demonstrates that there are no missing values in the data collection, allowing us to continue on to the next step.

```
pd.isnull(dataframe).sum()
```

```
squareMeters         0
numberOfRooms        0
hasYard              0
hasPool              0
floors               0
cityCode             0
cityPartRange        0
numPrevOwners        0
made                 0
isNewBuilt           0
hasStormProtector    0
basement             0
attic                0
garage               0
hasStorageRoom       0
hasGuestRoom         0
price                0
category             0
dtype: int64
```

Fig. 1. Null check of dataset

II. “Twitter US Airline Sentiment”

a. Data Selection:
Examine how travelers expressed themselves on Twitter in February 2015. Contributors were requested to identify good, negative, and neutral messages before classifying unfavorable causes using Twitter data from February 2015.
It has 3 columns and 14,640 rows as shown in Table II.

TABLE II
TWITTER US AIRLINE SENTIMENT

Column Name	Description
airline_sentiment	positive, negative, or neutral
text	text present in tweet
airline	Airline name
Target Variable	airline_sentiment

b. Data Pre-processing:
The data set’s description is shown in fig 2 below. It indicates that the data collection seems to have no missing values, allowing us to continue on to the next step.

```
pd.isnull(dataframe).sum()
```

```
airline_sentiment    0
airline              0
text                 0
dtype: int64
```

Fig. 2. Not null values

III. “Apartment rental offers in Germany”

a. Data Selection:
Rental offers taken from Germany’s largest online real estate website. The information was taken from https://www.immobilienscout24.de/, Germany’s largest real estate marketplace. Although immobilienscout24 features listings both for rental and sale properties, the data only comprises rental property offers.

It has 17 columns and 268,850 rows in Table III.

TABLE III
APARTMENT RENTAL OFFERS IN GERMANY

Column Name	Description
regio1	federal state of Germany
heatingType	Type of heating
newlyConst	is the building newly constructed?
balcony	does the object have a balcony?
yearConstructed	year of construction
hasKitchen	has a kitchen?
cellar	has a cellar
livingSpace	living space in sqm
condition	condition of the flat
interiorQual	interior quality
lift	is elevator available?
typeOfFlat	type of flat
geo_plz	ZIP code
noRooms	number of rooms
floor	which floor is the flat on
garden	has a garden?
baseRent	rent of the apartment
Target Variable	baseRent

b. Data pre-processing:
Data Set had lot of null values. Using drop function of pandas dropped column which had more than 60 percent null values. Fig. 3 decipts the code to remove columns.

```
for column in data_frame:
    if (data_frame[column].isnull().sum()/218354 * 100) > 60:
        data_frame.drop(column,axis=1,inplace = True)
```

Fig. 3. removing columns having more than 60 percent null values

The below Fig 4 shows the after replacing null values and removing unwanted columns.

```
pd.isnull(data_frame).sum()

newlyConst                        0
balcony                           0
hasKitchen                        0
cellar                            0
baseRent                          0
                                 ..
typeOfFlat_other                  0
typeOfFlat_penthouse              0
typeOfFlat_raised_ground_floor    0
typeOfFlat_roof_storey            0
typeOfFlat_terraced_flat          0
Length: 64, dtype: int64
```

Fig. 4. replacing null values and removing unwanted columns

B. Applied Machine Learning Algorithms

- Random Forest

  The Random Forest Classifier and Regression model is a machine learning approach for trying to resolve regression and classification problems. It uses the clear majority for classifications and the average for regression to generate decision trees from diverse datasets. Random Forests have the advantage of being able to handle data sets with both continuous and categorical, as in regression and classification. It outperforms the competition to categorization concerns. Forests are tree data structures with a node that allows them to extend in only one direction, thus the name. Figure 5 shows decision tree extensions that have had randomization applied to them, but in their own unique way. Random forest was chosen because of its versatility in terms of regression and classification. Next Overfitting is a serious problem that can undermine results, however the Random Forest technique's classifier will not overfit the model whether there are trees in the forest.

  Advantages of Random Forest is as follows:
  1. It is perfectly capable of handling missing values.
  2. In classification, Random Forest can reduce the risk of overfitting.

  Disadvantages of Random Forest is as follows:
  The most significant disadvantage of the Random Forest classifier is that it might produce delayed results if the number of trees is more than. Though this algorithm is quick to learn, it takes time to generate a forecast.

- KNN Classification

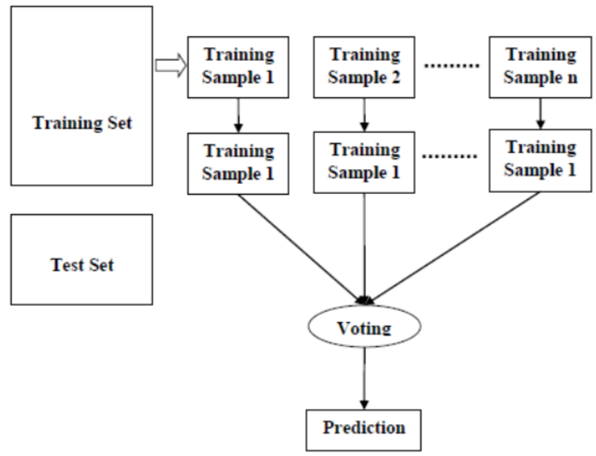

Fig. 5. Random Forest Working

  This algorithm is used to address categorization model problems. The K-nearest neighbor algorithm creates an imaginary border to categorize data. When new data points are received, the algorithm will try to predict them as near to the boundary line as possible. As a result, a larger k value means finer separation curves and simpler models. On the other side, smaller k values tend to overfit the data, resulting in complicated models.

  Advantages of KNN Classification is as follows:
  1. KNN is known as the Lazy Learner (Occasion based learning). It does not learn anything throughout the training period.
  2. It's simple to use.

  Disadvantages of KNN Classification is as follows:
  1. It is ineffective with huge data sets and dimensionality.
  2. Missing values, outliers, and noisy data make KNN extremely sensitive.

- Naive Bayes Classifier

  Naive Bayes classifier is a simple classifier based on Bayes' theorem and strong (naive) feature independence requirements. They are among the most basic Bayesian network models, but when used in conjunction with kernel density, they may attain high levels of accuracy.
  Advantages of Naive Bayes Classifier is as follows:
  1. It is simple and quick to predict the test data

set's class. It's also good at multi-class prediction.
2. When the requirement of independence is met, a Naive Bayes classifier outperforms alternative models such as logistic regression and requires less training data.

Disadvantages of Naive Bayes Classifier is as follows:
1. On the other hand, because naive Bayes is a lousy estimator, the probability outputs from predict probes should be regarded with caution.
2. The assumption of independent predictors is another flaw in Naive Bayes. In actual life, getting a collection of predictors that are totally independent is very impossible.

- Decision Tree Classifier

  Decision Tree is a Supervised Machine Learning Algorithm that makes judgments based on a set of rules, similar to how people do. A Machine Learning classification algorithm may be thought of as being created to make judgements. You may claim that the model predicts the class of a new, never-before-seen input, but the algorithm needs to select which class to assign behind the scenes.Some classification algorithms, such as Naive Bayes, are probabilistic, but there is also a rule-based method.
  Advantages of Decision Tree Classifier is as follows:
  1. can be quite helpful in resolving decision-making issues.
  2. It is beneficial to consider all of the possible solutions to an issue.
  Disadvantages of Decision Tree Classifier is as follows:
  1. The decision tree is complicated since it has several tiers.
  2. It may have an overfitting problem, which the Random Forest method can remedy.

- Linear Regression

  Linear regression is the most basic and extensively used approach of predictive modeling. Regression's purpose is to look at two things: Firstly, Can a group of regressors be used to anticipate an outcome (independent) variable? Secondly, Which factors are most highly predictive of the outcome variable, and how do they impact it (as indicated by the sign and magnitude of the beta estimates)?

  Advantages of Linear Regression is as follows:
  1. Linear regression is simple to use and analyze, and the output coefficients are straightforward.
  2. This strategy is the best to utilize when the relationship between the variables and dependent variables is linear since it is less difficult than other algorithms.

  Disadvantages of Linear Regression is as follows:
  1. Linear regression method The technique's boundaries are linear, therefore outliers can have a big influence on the regression.
  2. In linear regression, the independent and dependent variables are considered to have a linear relationship. That is, it is presumed that they are related in a straight line. It considers attributes to be separate from one another.

IV. Evaluation

The document explains how to run and apply the different algorithms, as well as their benefits and drawbacks. We found that in these types of decision tree classification, the Naive Bayes classifier, KNN classification, random forest, and linear regression are substantially more sensitive to outliers, therefore we'll take care of feeding the data to these approaches and generate a separate trained model for them. The following is a description of the algorithm and how it works:

A. *"Paris Housing Classification"*

For 75% and 25% of the data, we divided it into train and test data and used Random Forest Classifier, Decision Tree Regression, and KNN model to train data, achieving Accuracy Score values of 99.88 percent, 99.6 percent, and 99.6 percent, respectively, fig 6. shows f1 score for the model.

Decision tree is considered as best way for classification which can be seen in fig.7

Confusion matrix for Decision tree classification can be seen in fig.8
Confusion matrix for Random Forest classification can be seen in fig.9

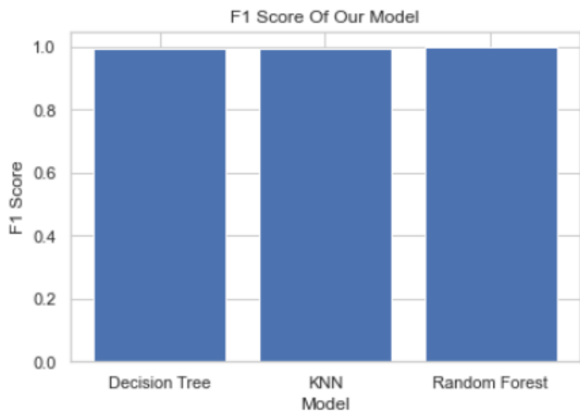

Fig. 6. f-1 score for model

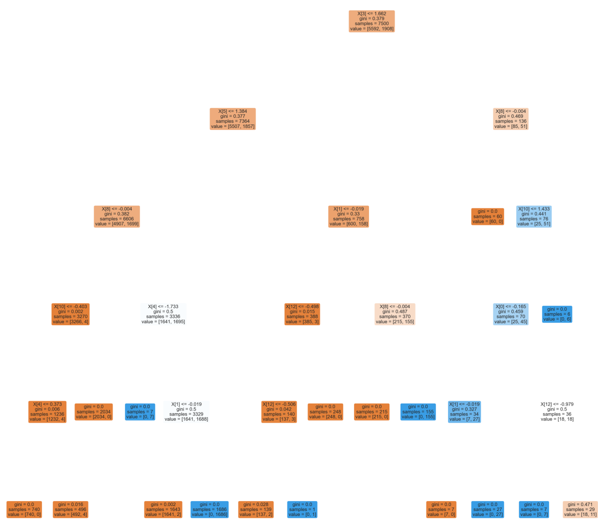

Fig. 7. Decision tree for model

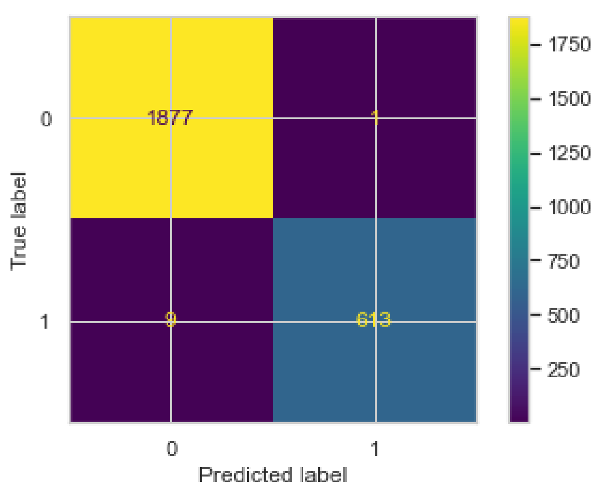

Fig. 8. Confusion matrix for Decision tree classification

Confusion matrix for KNN classification can be seen in

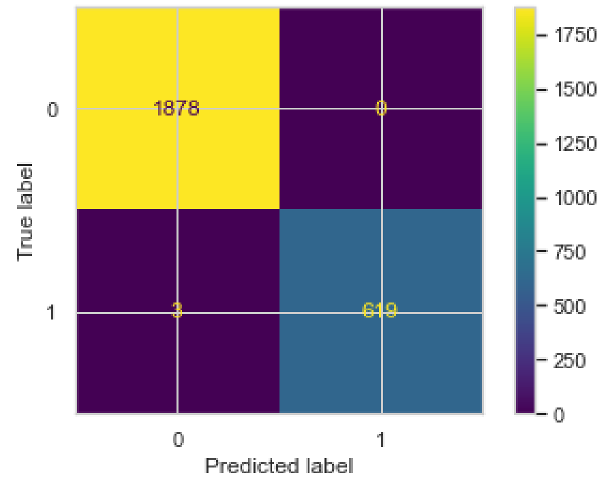

Fig. 9. Confusion matrix for Random Forest classification

fig.10

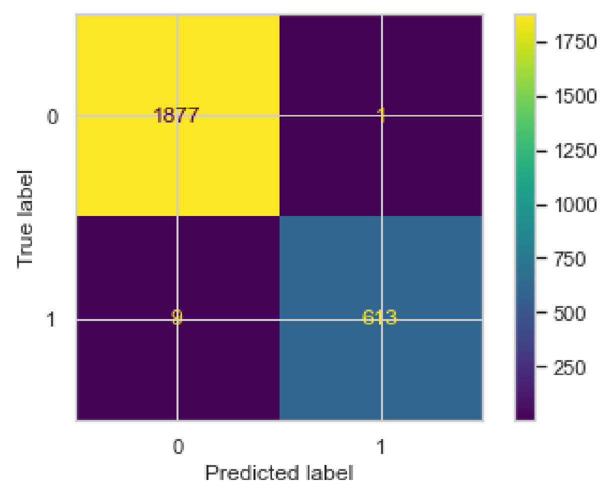

Fig. 10. Confusion matrix for KNN classification

B. "Twitter US Airline Sentiment"

I randomly split the data into train and test data for this data set and used a naive bayes classifier to get an accuracy score of 78.08 percent.As shown in fig 11, It also provides a confusion matrix to aid comprehension.

Figure 12 depicts the negative, positive, and neutral feelings for airlines in a bar chart. Consumer dissatisfaction with United Airlines is at an all-time high.

C. "Apartment rental offers in Germany"

I used Linear Regression and Random Forest Regression to forecast rent by partitioning data into

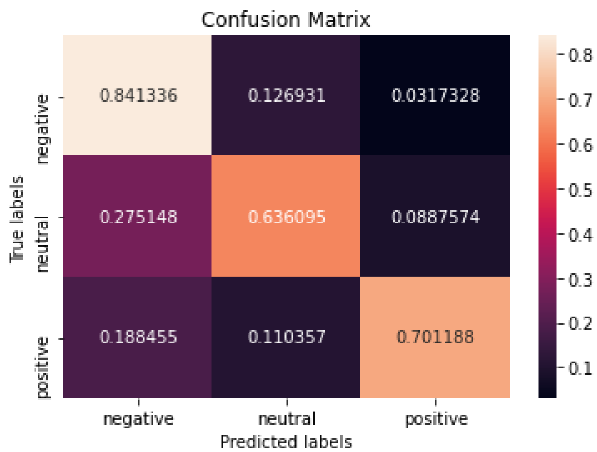

Fig. 11. confusion matrix of sentiments

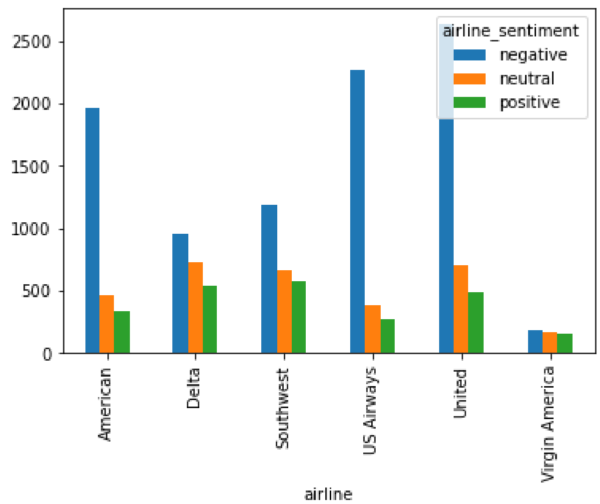

Fig. 12. Sentiments of US airline consumers

test and train groups of 80 percent and 20 percent, respectively. Between the two regression models, I have compared both model using MAE, MSE, and R2 score as shown in Table IV. Fig.13 shows Linear Regression and fig.14 shows Random Forest Regression.

D. Model Comparison

Following are the comparisons that were made after implementing six models (4 classification and 2 regression).

The following are the queries we choose to respond to before. The following are the findings from our investigation.

TABLE IV
ERROR VALUES

Regression Model	Mean Absolute Error	Mean Square Error	Root Mean Square Deviation
Linear Regression	0.19614	0.06350	0.80146
Random Forest Regression	0.14484	0.03808	0.88093

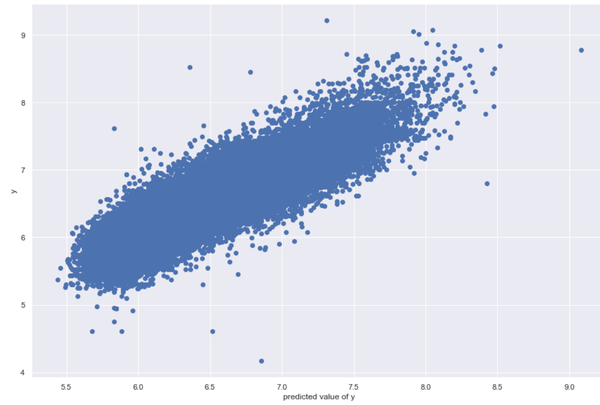

Fig. 13. Linear Regression

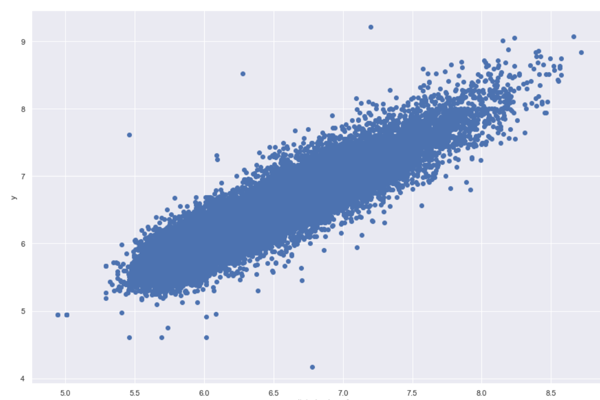

Fig. 14. Random Forest Regression

TABLE V
REGRESSION MODEL COMPARISON

Model	Result
Random Forest Regression	88.06
Linear Regresssion	80.14

TABLE VI
CLASSIFICATION MODEL COMPARISON

Model	Result
Decision Tree	99.60
Random Forest	99.88
KNN	99.60
Naive Bayes	78.08

1. When it comes to housing in Paris, there are two categories: basic and luxury. A basic property contains a storage place, a guest room, many former owners, a pool area, and a yard area. However, only a small percentage of premium homes have everything included.

2. During the sentiment study of US airline tweets, United Airlines received the most unfavorable attitudes, while Southwest Airlines received the most positive opinions. According to customer attitudes, while United's pricing is low, the services given are low-quality, whereas Southwest's price is high, the services offered are high-quality.

3. I utilized linear regression and random forest to forecast rent. The coefficients of accuracy are 81.06 and 88.06, respectively.

V. CONCLUSION

In this study, we used three data sets and performed a comparison analysis using six algorithms to provide an output that may be used in the future. To achieve the best execution result, we trained and tested our data in a variety of models and discovered that for regression models, Random Forest delivers 88.06 percent more accuracy than Linear, which provides 80.06 percent. When it comes to classification models, Random Forest comes out on top with a 99.88 percent accuracy rate (Table VI).

VI. FUTURE WORK

1. Graphical user interface will be added to the model using Python's Flask Framework, making it more interactive and understandable.

2. Our accuracy now ranges from 80 percent to 89 percent, which we will strive to increase in the future.

3. Developing a better prediction model: We'll develop a better predictor that can not only operate as a classification algorithm but also estimate a Paris home categorization based on demographic parameters.

4. In order to fulfill the aforementioned aim, we must first collect a dataset. To do so, we will either focus on collecting the dataset from multiple sources or create a survey to get the necessary data.

REFERENCES

[1] "A hybrid regression technique for house prices prediction."—Lu, Sifei, et al. — Industrial Engineering and Engineering Management(IEEM), 2017 IEEE International Conference on. IEEE, 2017— Conference Paper — Publisher: IEEE

[2] "Predicting house sale price using fuzzy logic and K-Nearest Neighbor." — Mukhlishin, Muhammad Fahmi, Ragil Saputra, and Adi Wibowo. — Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on. IEEE, 2017 — Conference Paper — Publisher: IEEE

[3] "House Price Forecasting using Data Mining Techniques" — Atharva chogle1, Priyanka khaire2, Akshata gaud3, Jinal Jain — International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, December 2017 — Conference Paper — Publisher: IEEE

[4] "Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java." — Febrita, Ruth Ema, et al. Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on. IEEE, 2017 — Conference Paper — Publisher: IEEE

[5] "Demonetization-Twitter Data Analysis using Big Data Hadoop" — Goyal et el. — IEEE, (2019), ISBN 978-1-5386-9346 — Conference Paper — Publisher: IEEE

[6] "Twitter Based Sentiment Analysis of GST —Prabhismaran Singh, Ravinder Singh Sawhney —(2019) Implementation by Indian Government" Spriger. — Conference Paper — Publisher: Spriger

[7] "Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble" — Fouad et el.— (Springer Nature 2018). — Conference Paper — Publisher: Spriger

[8] "Machine Learning-Based Sentiment Analysis for Twitter Accounts"(2018) — Ali Hasan, Sana Moin — Conference Paper

[9] "Real-Time Data Analysis Using Spark and Hadoop"—Khadija AZIZ, Dounia ZAIDOUNI —(2018 IEEE)— Conference Paper — Publisher: Spriger

[10] Cebula, Richard J.,"The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District," The Review of Regional Studies, Southern Regional Science Association, vol. 39(1), pp. 9-22, 2009.

[11] Alfiyatin, Adyan et al., "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia," International Journal of Advanced Computer Science and Applications, 2017, DOI: 10.14569/IJACSA.2017.081042.

[12] irong, Gu Zhu, Mingcang Jiang, Liuguangyan. "Housing price forecasting based on genetic algorithm and support vector machine." Expert Systems with Applications, vol. 38. pp. 3383-3386. 2011. DOI: 10.1016/j.eswa.2010.08.123.

[13] R. Hattori, K. Okamoto and A. Shibata, “Rent Prediction Models with Floor Plan Images,” IEEE 8th Global Conference on Consumer Electronics, Osaka, Japan, pp. 451-452, 2019. DOI: 10.1109/GCCE46687.2019.9015208.

[14] Sirmans, G. S., Benjamin, J. D. Determinants of market rent.Journal of Real Estate Research, Vol. 6(3), pages 357–380, 1991.

[15] Valente, J., Wu, S., Gelfand, A., Sirmans, C. Apartment rent prediction using spatial modelling. Journal of Real Estate Research, vol. 27(1), pp. 105–136, 2005.

[16] Zietz, E. N. Multifamily housing: A review of theory and evidence. Journal of Real Estate Research, vol. 25(2), pp. 185–244, 2003.

[17] Pow, Nissan, Emil Janulewicz, and L. Liu. 2014, Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal.

[18] Roland Füss, Jan A. Koller, “The role of spatial and temporal structure for residential rent predictions”,International Journal of Forecasting, Volume 32, Issue 4, 2016, pp. 1352-1368, Available: https://doi.org/10.1016/j.ijforecast.2016.06.001.

[19] Zainuddin et el. “Hybrid sentiment classification on twitter aspect-based sentiment analysis”(Springer 2017).

[20] https://www.kaggle.com/

[21] https://www.immobilienscout24.de/

[22] https://www.twitter.com/