Real Estate Price and Loan Approval Prediction

Sarthak Jain - 191IT145
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: sarthak94511@gmail.com

Yash Gupta - 191IT158
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: guptayash1104@gmail.com

Rishit - 191IT141
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: rishit.191it141@nitk.edu.in

***Abstract*—House is considered as one of the most valuable assets by many. Buying a dream house is a goal for most people. Two main hurdles in purchasing a home are: overpricing and the unavailability of a loan. Machine Learning Models can be used to predict the price based on its various features. Similarly, ML is used to predict the loan availability percentage. Integrating these features in house buying and selling website provides a one-stop solution for buyers and sellers, to browse properties, add their properties for selling, predicting price and checking for loan availability.**

I. Introduction

In this developing global village with the increasing human activities such as industrialization, the land has become a very highly-priced resource. Owning a dream house is a dream of almost every person. People spend their life's savings to get a roof over their heads and consider their home their greatest asset. It is also a precious investment tool. The real estate market is primarily increasing and provides a decent return over a period of time. These reasons make the real estate market always busy and moving, but individuals have many hurdles while buying/selling a house. Some of those hurdles are finding a convenient house, bargaining, commission of middleman involved, estimating accurate price, and checking for loan availability.

A lot of time is wasted in searching for an ideal buyer/seller. Property dealers help in finding buyers and sellers, but they also charge hefty commissions from both parties. A lot of time, resources are spent on travelling to various parts of the city to select their dream house, but most of them are rejected by the buyer as it doesn't suit his needs and requirements. The same story is with sellers. They struggle to find genuine buyers who will pay a fair price for the property.

The soaring land prices and intermediaries involved trouble both the buyers and sellers as these mediators charge heavy commission on the deals they initiated. So there is the need for a platform where users can easily find their ideal home, and sellers can advertise and list their properties to find a buyer fast instead of paying heavily to advertise and giving money to mediators to sell their house. There are several such websites and apps available. Still, they don't solve other significant issues involved, which are predicting a fair price for their home and getting the surety that they will be able to avail a loan for that amount. This project solves these stated issues using Machine Learning Models and probabilistic analysis. Several parameters like city, the locality in town, size of the house, BHK, and the type of house are considered to predict the price of the house. But a suitable dataset is required to train the model to make accurate predictions. Due to the unavailability of a proper dataset, the dataset was generated after scrapping the data from makaan.com, a leading house listing platform. The cities considered are Bangalore, Agra, Ahmedabad, Amritsar, Delhi, Hyderabad, Lucknow, Chennai, Mumbai and Noida. The major localities in these cities are considered so that the project can be helpful to a large section of the population. Similarly, several parameters are considered for predicting loans, such as the applicant's monthly income, education, dependents, etc. The Machine Learning Models are coupled with a web app to provide a one-stop solution to the users. The web app provides the feature to browse properties, add a property, predict property price and loan availability. The user will only require to enter his details for loan only once and then whenever the price will be predicted, loan availability status will also be showed to the user.

II. Problem Statement

Predicting Loan availability and real estate price using Machine Learning.

A. Objectives

- Scrapping Data and cleaning it to create property price dataset for major Indian cities
- Building Machine Learning Model for predicting the property price based on various features
- Building Machine Learning Model for predicting Loan Availability
- Building web app (front-end and back-end) to provide a real life use to the work done.

III. Literature Survey

Through the paper [1] by W.T. Lim and Wang and article [2] we learnt about various ways in which regressions can be applied for housing and real estate data. By paper [3] we learnt about ways of performing statistical analysis on dataset involving real estate data. We also learnt about ways of encoding categorical data. Through article[5] we learnt about ensembled based learning and also about advanced regression and classification algorithm extreme gradient boost (X.G Boost).

By referring paper[6] by James, Witten, Hastie and Tibshirani, we learnt how correlation can be used to analyse datasets. By referring paper [7] by A. Azadeh, B. Ziaei, and M. Moghaddam we learnt about various optimastions and about applying multiple algorithms and picking best out of them for use.For learning more on data cleaning we reffered to Data Cleaning in Python: the Ultimate Guide (2020) by Lianne since there were many faulty entries in our dataset. For learning more on React framework we reffered to its official documentation. For Back-end, in order to learn more on flask we referred to international journal on efficient way of development using python and flask by Mr.Prakash P Lokande. We also reffered to the official documentation of flask.

IV. Methodology

A. Property Price Prediction

1) Dataset: There is no dataset available which has data for Indian real estate properties. Therefore, we scrapped a website called makaan.com to obtain the data for ten Indian cities: Banglore, Agra, Noida, Delhi, Hyderabad, Amritsar, and Chennai Mumbai, Lucknow, Ahemdabad. The dataset contains features like the name of the locality in a particular city, price of real estate as listed by the seller, area of properties in square feet, BHK, type of real estate whether the property is an apartment, villa, independent house, studio apartment, residential plot. Since data is scrapped city-wise, the city column is inserted in the dataset, which contains the name of the city to which real estate property belongs. The dataset contains 40000 tuples, each representing unique real estate property. There were few tuples with the value corresponding to null, so they are replaced with the median of remaining values of bhk.

In order to apply various regression algorithms on the dataset, we need to encode values corresponding to city names, the name of the locality in the city, and the type of property. In order to perform the encoding, label encoder is used. It has a function called fit transform to perform encoding. Each city is given a unique code. It also has a function called inverse transform to obtain city corresponding to code so that a dictionary can be created with city name as key, which can be later used to make predictions. A similar process is followed for name of locality and type of real estate property. Three dictionaries obtained by the following above mentioned are converted to CSV files which will be used to find the encoding of input data given by user to for finding prediction.

2) Machine Learning Algorithms: The dataset after performing encoding is trained on multiple regression algorithm. First algorithm used for performing regression is decision tree regressor. Max depth of decision tree is varied to find best performing decision tree. The other regression algorithm used are linear regressor, lessor regressor, ridge regressor. To improve our results advanced ensemble based regression algorithms like extreme gradient boost (xgboost) regressor, cat boost regressor, light gbm regressor are used. GridSearchCV is used to find the best model and best parameter. Depth, learning rate, number of estimators are among the parameters which are varied in GridSearchCV to obtain the best regression results for the obtained dataset. GridSearchCv finds the best model based on the cross-validation score obtained by splitting data into test train dataset for permutation of parameters for a particular algorithm. A pickle file corresponding to best performing dataset is obtained so that it can be used in the backend of the application, which will help in performing regression on data given as input by the user.

B. Loan Approval On Predicted Price

We also look to provide the user information regarding his ability to buy the property depending on certain factors which are queried by the user. The proposed model focuses on forecasting a customer's ability to repay a loan by evaluating their behaviour. The collected customer behaviour serves as the model's input. The classifier's output can be used to determine whether the customer request should be approved or rejected. Loan prediction and severity can be projected using various data analytics technologies. In this procedure, data must be taught using various algorithms, and then user data must be compared to trained data in order to forecast the type of the loan.

To extract patterns from a publicly available loan approval dataset and then develop a model based on those patterns.The training data set is now provided to the machine learning model, and the model is trained using this data set. Every new applicant's information entered on the application form serves as a test data set. Following the testing procedure, the model predicts if the new applicant is a good candidate for loan approval based on the inferences drawn from the training data sets.

1) Dataset: The dataset contains 129 features out of which 9 features are extracted using feature engineering. The features are plotted against the Loan Status which is the target variable in our dataset. The bar plots are then analysed to find out the impact that these features create in determining the Loan Status that is it's status of being approved. It is observed that 9 features turn out to be very impactful in determining our target variable. These features include - number of dependants, gender, level of Education, the type of residence, annual income, loan Amount and the number of bedrooms, halls and kitchens. So in order to decide the impact of each of these features, we performed feature engineering on our refined dataset. We have plotted these

features against our target i.e, Loan Status whose results are as shown below in order to draw conclusions regarding their weights in the machine learning model.

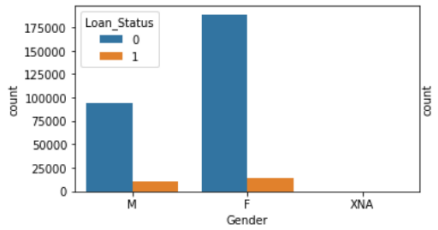

Fig. 1. Gender vs Loan Status

From the graph we can conclude that males have a higher ratio of the loan being approved as compared to the other classes that is females.

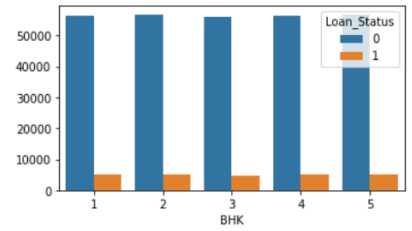

Fig. 2. BHK vs Loan Status

The number of bathrooms,halls and kitchens do not have any significant impact in determining the Loan status as can be seen from the graph.

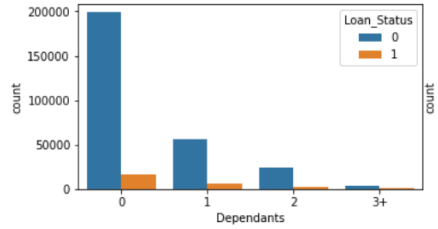

Fig. 3. Number of Dependants vs Loan Status

The lesser the number of dependants the higher is the chance of the loan being approved.

The higher the education one has achieved in their life the higher is their probability of loan being approved.

The housing doesn't definitely determine the status of loan being approved as their distribution is uniform across all classes.

The married couples have a slightly higher chance compared to the singles and the other classes have very less probability of the loan being approved.

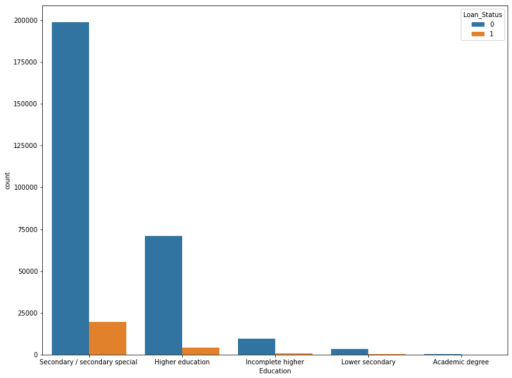

Fig. 4. Education vs Loan Status

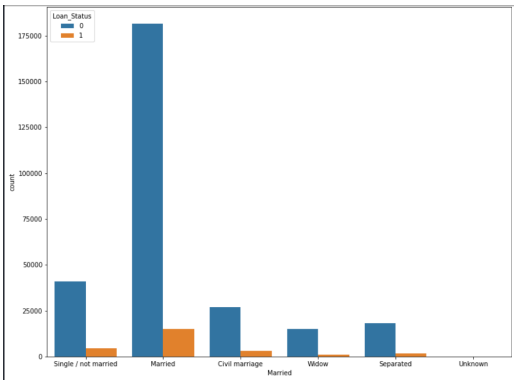

Fig. 5. Married vs Loan Status

The dataset is then refined to these 9 features and data munging is performed on the accumulated data. Firstly the null values in the data are filled such that the categorical features are filled with the mode values and the numerical missing values are filled with the mean of the data of that feature. The dataset is then checked for any outliers which can be removed to increase the efficiency of the machine learning models. Once the data is cleaned, certain categorical variables are not in their numerical form due to which these features need to be encoded so that the models can train and predict values using these features. In order to perform one hot encoding the label encoder library of python is used and the features are then encoded.

2) Machine Learning Algorithms: The dataset is then made ready for training and testing. In order to train and test the model, we have made use of cross validation scores with 5 cross validation folds. The models used to perform the training are Logistic Regression, GaussianNB, MultinomialNB, Decision Tree classifier and RandomForstClassifier. In order to further evaluate the performance of our ML model, we have applied GridSearchCV with the above models and performed HyperParameter Tuning on all of the above mentioned Models. They are then compared based on their cross validation scores and the best model is then picked for pickle file generation. The model is then dumped into a pickle file which is then used on the server to make predictions.

In order to understand the model selected in much more detail, we have plotted the confusion Matrix and observed the count false positives and the false negatives, which turned to be quite low as the accuracy of our model was quite high. We further found out the specificity and coercivity in order to draw conclusions that are model very well fit the dataset and made very accurate predictions.

Once the analysis of our model was complete and the model finally selected the pickle file was loaded on to the web server to make predictions on user entered values at the front-end.

V. Results and Analysis

	model	best_score	best_params
0	decision_tree	0.975019	{'criterion': 'mse', 'max_depth': 9, 'splitter...
1	ada	0.656284	{'learning_rate': 1, 'n_estimators': 200}
2	xgboost	0.913266	{'max_depth': 15, 'n_estimators': 15}
3	lightGBM	0.978750	{'learning_rate': 0.09, 'max_depth': 4}
4	catboost	0.875169	{'depth': 3}

Fig. 6. GridsearchCV Hyperparameter tuning

LightBGM has the highest accuracy score of 97.87 so it has been selected as our model for regression.

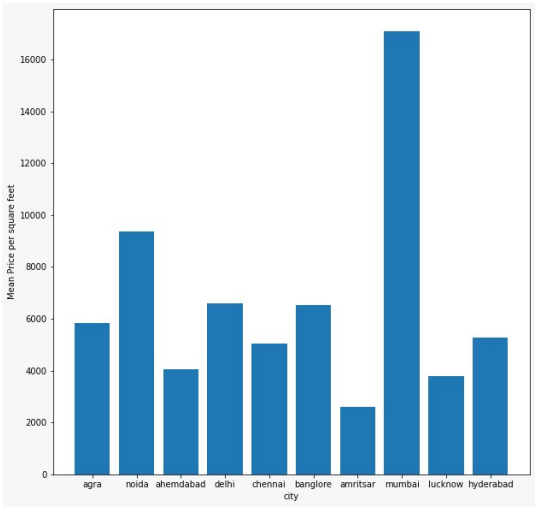

Fig. 7. Average price per square feet of various cities

The average price per square feet has been plotted against the cities which shows that Mumbai is the costliest city to live in.

Correlation matrix shows the correlation between the features and the price predicted and how they impact the prediction.

The histogram of frequency vs square feet has been plotted for the dataset to know the distribution of data points in the dataset.

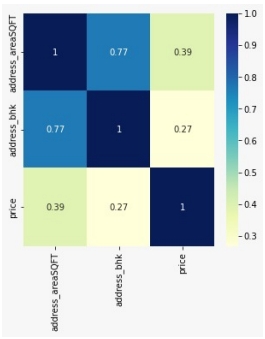

Fig. 8. Correlation Matrix

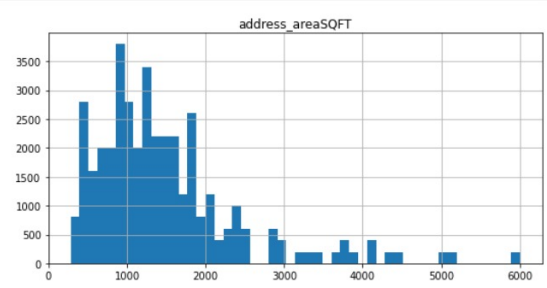

Fig. 9. Frequency vs Square Feet

The barplot of bhk vs frequency has been plotted for the dataset to know the distribution of data points in the dataset.

From the table it is clearly visible that logistic regression and naive bayes gaussian have the highest accuracy amongst all algorithms when gridsearchCV was applied on the mentioned algorithms in the table. However the cross validation score of Logistic Regression was slightly higher due to which the logistic regression model was chosen as our go-to model.

As our model has an accuracy rate of 92 percent the number of true positives and true negatives outnumbers the number of false positives and negatives.

The values of sensitivity and specificity are calculated on the formulas as mentioned :

sensitivity = TP/(TP +FN)

specificity = TN/(FP +TN)

The ML models are integrated with a web app for easier use. Other functionalities such as browsing properties available for sale, adding property for selling are also provided.

VI. Conclusion

This project aims to perform statistical analysis on real estate data of various cities in India. Since real estate datasets are not available for free, so in order to obtain dataset we scrapped site called makaan.com for ten cities. We obtained features like area in square feet, BHK, locality in which property is located, type of property like whether its villa, apartment, residential etc, price of property. The dataset which we obtained has a large number of actual data which helped us in performing various statistical operations. In our application

Fig. 10. bhk vs frequency

	model	best_score	best_params
0	logistic_regression	0.919271	{'C': 1}
1	naive_bayes_gaussian	0.913847	{'var_smoothing': 1e-12}
2	naive_bayes_multinomial	0.481845	{}
3	decision_tree	0.841606	{'criterion': 'entropy', 'splitter': 'best'}

Fig. 11. GridsearchCV applied on the machine learning models to perform hyper-parameter tuning for loan approval

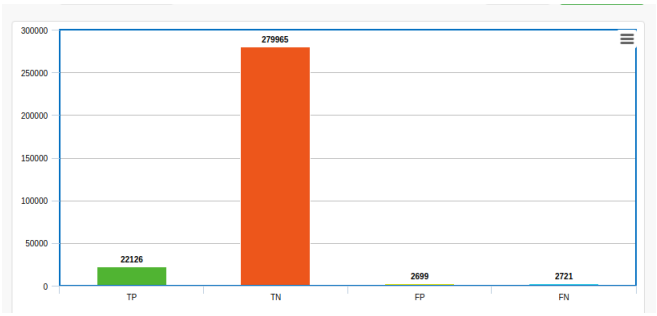

Fig. 12. Comparison of true positives and negatives with false positives and negatives

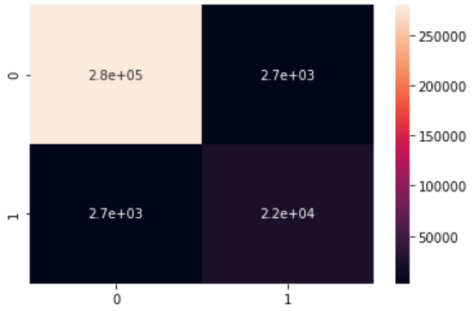

Fig. 13. Confusion matrix

which has two parts one is prediction of price of real estate data and other is whether an user is capable of getting loan based on various factors.

A separate dataset which connects to the scrapped data based on common columns like bhk, area in square feet, price is used whereas other details like education, income of applicant is taken while he/she runs the application for the first time so that he need not to enter details for properties multiple times. This loan prediction allows us to get the probability of getting loan. For prediction of real estate price dataset obtained by scrapping of makaan.com site is trained on multiple regression algorithms like decision tree regressor, xgboost regressor, catboost regressor etc and pickle file is generated for best performing model and used in the backend of the application whereas for prediction of eligibility of getting loan and for getting probability of getting loan dataset is trained on various classification models like decision tree classifier, naive bias classifier, xgboost classifier, catboost classifier etc. For getting probability of getting loan a pickle file is generated which is used in backend. As far as future work is concerned, currently this application works for 10 cities which can be extended for more national and international locations. Application can be hosted for public use and the dataset scrapped can be made public for general use. Also a property review system can be added where verified users can rate property and can write about its pros and cons, about the locality etc. Since there were sufficient details in the dataset that we scraped, we were able to perform analysis on data like finding expectation of price per square feet in various cities,finding specificity etc on loan part of the project.

For banking institutions, loan acceptance is a critical step. The loan applications were either approved or rejected by the system. Loan recovery is a significant contributing factor in a bank's financial statements. It is quite difficult to forecast if the customer will be able to repay the loan. Many researchers have been working on loan approval prediction algorithms in recent years. Machine Learning (ML) approaches are extremely beneficial for predicting outcomes when dealing with enormous amounts of data. Multiple machine learning techniques are used in this work to predict customer loan approval.We also look to provide the user information regarding his ability to buy the property depending on certain factors which are queried by the user. These features are recorded and their ability to repay the specified amount is displayed. In order to arrive at the decision of loan approval we have considered certain features whose trends are recorded as described below.

From the graph of gender vs loan status we can conclude that males have a higher ratio of the loan being approved as compared to the other classes that is females. The number of bathrooms, halls and kitchens do not have any significant impact in determining the Loan status as can be seen from the graph. The lesser the number of dependants the higher is the chance of the loan being approved. The higher the education one has achieved in their life the higher is their probability of loan being approved. The housing doesn't definitely determine the status of loan being approved as their distribution is uniform across all classes. The married couples have a slightly higher chance compared to the singles and the other classes have very less probability of the loan being approved. So it is clearly visible that the loan approval soleley is not determined by the income of an individual rather it depends on a lot of other factors as mentioned above.

```
The specificity is :  0.9904515608637817
The sensitivity is :  0.8904897975610738
```

Fig. 14. Specificity and sensitivity values

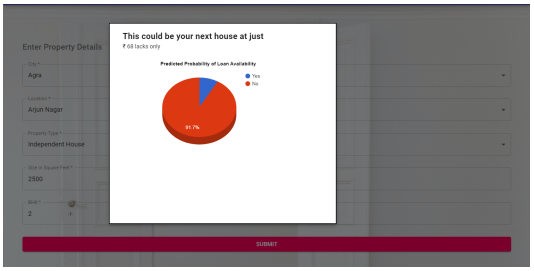

Fig. 15. Price and Loan availability prediction.

ACKNOWLEDGMENT

We would like to express our gratitude to Mr.Anand Kumar for his assistance in making our project successful, along with the IT department, NITK, for providing us the opportunity to work on this project.

REFERENCES

[1] W. T. Lim, L. Wang, and Y. Wang, —Singapore Housing Price Prediction Using Neural Networks,Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., vol. 12, pp. 518–522, 2016

[2] Mathematical/ Analytical Modelling and Comuter Simulation, 2010, no.1

[3] V. Limsombunchai, —House price prediction: Hedonic price model vs. artificial neural network, Am. J., 2004.

[4] M. Risdal, "Predicting House Prices Playground Competition: Winning Kernels", 2017. [Online]. Available: http://blog.kaggle.com/2017/03/29/predicting-house-prices-playgroundcompetition-winning-kernels/ [Accessed: 25-Mar-2019].

[5] A. Nair, "How To Use XGBoost To Predict Housing Prices In Bengaluru: A Practical Guide", 2019. [Online]. Available: https://www.analyticsindiamag.com/how-to-use-xgboost-to-predicthousing-prices-in-bengaluru-a-practical-guide/ [Accessed: 25-Mar-2019].

[6] James, Witten, Hastie and Tibshirani, 2013. An Introduction to Statistica Learning with Applications in R,Springer-Verlag NewYork.

[7] A. Azadeh, B. Ziaei, and M. Moghaddam, —A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations, Expert Syst. Appl., vol. 39, no. 1, pp. 298– 315, 2012

[8] https://www.sas.com/en/_us/insights/analytics/machine-learning.html

[9] De Cook, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." Journal of Statistics Education, vol. 19, no. 3, 2011.