**Credit Risk Assessment and Real Estate Price Prediction using Machine Learning**

Contents

[Introduction 3](#_Toc144721260)

[Problem Definition 3](#_Toc144721261)

[Business Applications of Classification and Regression Algorithms 4](#_Toc144721262)

[Credit Risk Assessment 4](#_Toc144721263)

[House Price Prediction 4](#_Toc144721264)

[Leveraging Data-Driven Insights for Informed Business Decisions 5](#_Toc144721265)

[Analysing Datasets 5](#_Toc144721266)

[Credit Score 5](#_Toc144721267)

[Data Cleaning: Handling Missing Values(Text Data) 5](#_Toc144721268)

[House Price Prediction 6](#_Toc144721269)

[Data Cleaning: Handling Missing Values(Numerical Data) 6](#_Toc144721270)

[Enhancing Decision-Making with Data-Driven Solutions 6](#_Toc144721271)

[Steps Towards Deploying Our Data-Driven Solution 7](#_Toc144721272)

[Versatile Data Analysis: Classification and Regression Challenges Solved 8](#_Toc144721273)

[Classification and Regression Task Variables 8](#_Toc144721274)

[Training Data Composition 9](#_Toc144721275)

[Evaluating Model Performance 9](#_Toc144721276)

[Conclusion 9](#_Toc144721277)

[References 10](#_Toc144721278)

# Introduction

In this data-driven decision-making journey, I have taken on two crucial challenges: classifying credit scores and predicting house prices. Each of these tasks carries tremendous significance within its respective domain. The accurate classification of credit scores is paramount for financial institutions, as it directly impacts lending decisions and risk assessments. On the other hand, predicting house prices holds the key to empowering real estate professionals and investors, facilitating well-informed choices and streamlining transactions.

To support these endeavors effectively, I have carefully curated relevant datasets from reputable sources like Kaggle. These datasets encompass a wide range of features, both numerical and categorical, providing a solid foundation for the development of machine learning models. My overarching goal is to enhance decision-making processes, reduce risks, and usher in improvements in financial and real estate management.

# Problem Definition

The issue at hand concerns the classification of credit scores and the prediction of house prices, two essential components of data-driven decision-making. These activities address various but fundamental problems, each of which is very significant in its own field.

The goal is to create reliable machine learning models that can properly classify people into various credit score categories. The financial stability and creditworthiness of an individual are strongly correlated with their credit score. When deciding whether to lend money, financial organisations, including banks and credit bureaus, mainly depend on credit ratings. This issue is critical because it affects people's ability to obtain credit, the conditions of loans, and financial possibilities. In this situation, a solution would have enormous advantages. The technology can assist financial institutions in making educated lending decisions, lower the risk of defaults, and perhaps increase access to credit for deserving people by automating and improving the credit assessment process. Additionally, it improves uniformity, simplifies processes, and eventually aids in better financial management for both people and organisations.

Here, the goal is to create prediction algorithms that can calculate house selling values based on numerous property characteristics. For buyers, sellers, investors, and real estate professionals, correct property assessment is essential in the real estate industry. It aids customers in making wise purchasing selections and helps sellers set competitive listing prices. It also enables investors to spot profitable investment possibilities. This issue is significant because it has a direct bearing on real estate deals, investment plans, and financial security. A strong solution gives real estate businesses a competitive edge, makes transactions go more smoothly, simplifies the negotiation process, and boosts investment returns on real estate.

# Business Applications of Classification and Regression Algorithms

The choice of regression algorithms (Linear Regression and Random Forest) for house price prediction and classification algorithms (Random Forest and Support Vector Classifier, SVC) for credit score classification suggests that the primary objectives of this solution are to address two particular business problems:

## Credit Risk Assessment

Business choice: This business choice is probably connected to determining a person's or an entity's creditworthiness. When deciding whether to accept loan applications, issue credit cards, or extend lines of credit, financial organisations like banks and credit agencies must classify credit scores.

Use Case: By predicting an applicant's credit score category based on several characteristics including income, job history, financial behaviour, and more, this system may assist in automating the credit evaluation process. This forecast helps informing judgements about whether to extend credit and under what conditions (Safiya Parvin, A. and Saleena, B. (2020)).

Benefits: Businesses may improve their credit risk assessment process, increase decision consistency, minimise manual burden, and perhaps lessen default risk by applying classification algorithms like Random Forest and SVC.

## House Price Prediction

Business Decision: The real estate sector or property market is relevant to the business decision that is being supported in this instance. For property purchasers, sellers, investors, and real estate professionals to make educated judgements regarding real estate transactions, accurate home price forecast is crucial.

Use Case: Using information on the size, location, number of bedrooms, amenities, and other characteristics of a property, this solution seeks to forecast the selling price of homes. Real estate buyers and sellers may use it to gain important insights about how to properly establish listing pricing and negotiate().

Benefits: Companies in the real estate market may provide more precise property values, optimise pricing strategies, and allow smoother real estate transactions by applying regression algorithms like Linear Regression and Random Forest.

# Leveraging Data-Driven Insights for Informed Business Decisions

The solution consists of Credit Score Classification and House Price Prediction models, which utilize data-driven insights. Machine learning algorithms like Random Forest and Support Vector Classifier improve financial institutions' credit risk analysis, enabling intelligent decisions about loan approvals, credit card issuing, and credit line extensions. This data-driven strategy encourages cautious lending practices, lowers default rates, and increases profitability. Linear Regression and Random Forest models provide crucial insights for property purchasers, sellers, investors, and real estate professionals in house price prediction. This data-driven strategy enhances decision-making in the volatile real estate market, leading to more effective transactions and greater returns for investors and property owners. Incorporating data-driven accuracy into decision-making processes provides significant value, improving efficiency, accuracy, and consistency of credit evaluations and property price forecasts, lowering risk and increasing corporate profitability. Data-driven solutions support wiser, more trustworthy decision-making, unleashing value across the financial and real estate industries.

# Analysing Datasets

## Credit Score

The data for Credit Score Classification is taken from the Kaggle dataset labelled "Credit Score Classification Dataset." With "Credit Score" as the goal variable, this dataset contains characteristics like Age, Gender, Income, Education, Marital Status, Number of Children, and Home Ownership. We can evaluate a person's creditworthiness based on their demographic and financial features according to the dataset's characteristics, which include a combination of numerical and category information(Mandala, S. K. (2023)).

Link to Dataset: <https://www.kaggle.com/datasets/sujithmandala/credit-score-classification-dataset>

### Data Cleaning: Handling Missing Values(Text Data)

Handling missing values in text data is crucial for data preprocessing, ensuring the dataset's completeness and quality. This task addresses missing education information in the 'Education' column of the Credit Score Classification Dataset. To identify missing values, the dataset is loaded and checked for NaN or empty strings. Some applicants have missing education details due to incomplete data entry or omitted information. To handle these values, a common strategy is to fill them with a suitable placeholder value, such as 'Unknown'. A second check is performed to confirm the removal of missing values. This ensures consistency and reliability for machine learning tasks. Addressing missing education information for credit applicants improves the dataset's completeness and suitability for credit score classification analysis (Houari, R. *et al.* (2014)).

## House Price Prediction

The "Housing Price Prediction" Kaggle dataset serves as the source of the information for "House Price Prediction." This dataset includes information on the size of the property, the number of bedrooms, baths, and storeys, as well as amenities (such as parking and air conditioning). "price," which represents the selling price of homes, is the goal variable in this dataset. The dataset includes a variety of variable types, such as numerical and binary categorical variables, which enables us to investigate the effects of various property qualities on housing prices(KUMARdatalab, H. (2023)).

Link to Dataset: <https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction>

### Data Cleaning: Handling Missing Values(Numerical Data)

Addressing missing values in numerical data is crucial for accurate analysis and modeling, especially when dealing with incomplete or inconsistent data. In this task, we focus on the 'bathrooms' column in the Housing Price Prediction Dataset. We load the dataset and check for missing values using the isnull().sum() method. Upon inspection, we find some properties have missing bathroom information due to data entry errors or incomplete records. To address this issue, we fill missing values with an appropriate estimate, such as the median value of the 'bathrooms' column. This approach maintains the data's distribution and ensures representativeness. After filling missing values with the median, we perform a second check to confirm no more missing values in the 'bathrooms' column, validating the data cleaning task and ensuring the dataset's suitability for further analysis and modelling(Dubey, A. and Rasool, A. (2019)).

Both datasets have been used to actual data analysis situations and are real. They provide useful information for developing, training, and assessing machine learning models that handle the particular business issues mentioned above. These databases' insights can improve decision-making, lower risks, and aid in real estate and financial management for both people and institutions.

# Enhancing Decision-Making with Data-Driven Solutions

The solution consists of Credit Score Classification and House Price Prediction models, which offer significant value by utilizing data-driven insights. Machine learning algorithms like Random Forest and Support Vector Classifier (SVC) can significantly improve financial institutions' credit risk analysis, allowing them to make intelligent decisions about loan approvals, credit card issuing, and credit line extensions. This data-driven strategy encourages cautious lending practices, lowers default rates, and increases profitability (Torkjazi, M. and Raz, A. K. (2023)).

Linear Regression and Random Forest models provide crucial insights for property purchasers, sellers, investors, and real estate professionals in house price prediction. These algorithms estimate property prices based on various property characteristics, helping stakeholders find attractive investment opportunities, optimize pricing tactics, and make data-informed decisions. This data-driven strategy could significantly enhance decision-making in the volatile real estate market, leading to more effective transactions and greater returns for investors and property owners(Jaiswal, J. K. and Samikannu, R. (2017)).

In conclusion, incorporating data-driven accuracy into decision-making processes is poised to provide significant value. Automation and optimization will improve the efficiency, accuracy, and consistency of credit evaluations and property price forecasts, ultimately lowering risk and increasing corporate profitability. Additionally, these models will enable knowledgeable financial and real estate decisions, leading to a more informed and assured customer base. Overall, data-driven solutions can support wiser, more trustworthy decision-making, unleashing value across the financial and real estate industries.

# Steps Towards Deploying Our Data-Driven Solution

Our data-driven solution's implementation, which includes Credit Score Classification and House Price Prediction models, entails a number of clearly defined procedures to guarantee that it is successfully integrated into actual company operations.

Data acquiring and Preparation: The first phase is acquiring and getting ready the necessary data. We'll make sure the data sources are available and current. To prepare the data for modelling, further data pretreatment operations including cleaning, feature engineering, and transformation will be carried out (Sagi, S. L. N. *et al.* (2021)).

Model Development and Evaluation: Using the provided datasets, machine learning models such as Random Forest, Support Vector Classifier (SVC), Linear Regression, and Random Forest will be created. To guarantee their correctness and robustness, these models will go through a thorough evaluation process employing methods like cross-validation, hyperparameter tweaking, and performance metrics assessment (Pathak, S., et al., (2018)).

After developing a model, the best models will be chosen and validated based on their effectiveness and appropriateness for the intended use cases. Model validation guarantees that the selected models provide accurate predictions and generalise effectively to new data.

In summary, the deployment of our data-driven solution is a methodical procedure that includes model creation and data preparation. This strategy guarantees that our solution will successfully answer company needs and continue to create value over time.

# Versatile Data Analysis: Classification and Regression Challenges Solved

I have used Machine learning models to solve two main data analysis tasks: credit score classification and house price prediction. The credit score classification task utilized models like Random Forest and Support Vector Classifier (SVC) for accurate categorization of individuals into creditworthy and non-creditworthy groups in credit risk assessment. The house price prediction task utilized regression models like Linear Regression and Random Forest Regression, which capture complex relationships and handle non-linearities. Combining both tasks demonstrated the versatility of machine learning in addressing diverse data analysis needs. Classification models proved invaluable for categorizing data into distinct classes, while regression models demonstrated their ability to predict numerical values and estimate quantities. Overall, these tasks showcased the broad applicability of machine learning in tackling real-world challenges across diverse domains.

# Classification and Regression Task Variables

With "Credit Score" as the objective variable, the main goal of the classification job for credit score classification was to forecast people's credit scores. The following factors were taken into account while making this prediction: "Age," "Gender," "Income," "Education," "Marital Status," "Number of Children," and "Home Ownership." These characteristics were essential in determining a person's creditworthiness since they provide information about their financial, lifestyle, and demographic characteristics.

On the other hand, the objective of the regression job for house price prediction was to predict the selling values of homes. The goal variable in this case was "price," which stands for the selling price of a property in terms of money. With the exception of "price," we used all other columns in the dataset as characteristics to make this prediction. These qualities included a wide variety of property details, including size, number of rooms, baths, storeys, and extras like parking and air conditioning. These characteristics helped us identify the variables affecting property prices and make precise predictions about their values.

In conclusion, each of these two projects required a unique collection of characteristics that were customised to the task at hand. The regression work intended to estimate housing values by taking into account different property variables, whereas the classification task sought to categorise people based on their credit ratings.

# Training Data Composition

The training data would be a subset of the corresponding datasets for the classification problem of "Credit Score Classification" and the regression task of "House Price Prediction." The training data for the Credit Score Classification job would be a subset of the credit score dataset that included the requested characteristics (such as "Age," "Gender," "Income," etc.) and the relevant target variable, "Credit Score." The classification models, such as Random Forest and Support Vector Classifier (SVC), would be trained using this subset in order to discover the underlying patterns and connections between the features and credit scores.

The training data for the regression task for predicting house prices would also come from the dataset for that task. All columns of this training dataset would be present, excluding the target variable "price." To learn how different property variables affect the selling prices of homes, it would be used to train regression models like Linear Regression and Random Forest Regression.

The basis for model learning and parameter optimisation in both situations is the training data. When faced with fresh, unexplored data, the models use what they have learned from existing data to create precise predictions or classifications. Building reliable and efficient machine learning models that are suited to each unique job requires access to training data.

# Evaluating Model Performance

To evaluate machine learning models' performance in credit score classification and house price prediction tasks, a range of evaluation metrics and techniques are employed. Accuracy, precision, recall, F1-score, are crucial for evaluating the model's ability to identify true positive cases and minimize false positives (Alam, S. *et al.* (2022)). In the regression task, metrics like MSE and R-squared are used to assess the average squared difference between predicted and actual house prices (Nazarov, F. M. and Yarmatov, S. (2023)). Cross-validation techniques are employed to ensure robustness and prevent overfitting. The choice of metrics and techniques depends on the specific goals and requirements of each task, with the goal being to select models that demonstrate high accuracy and reliability in their respective domains.

# Conclusion

Finally, I consider the classification of credit scores and the prediction of house prices to be two critical issues in our data-driven decision-making journey. These tasks hold immense significance in their respective domains.

In the realm of Credit Score Classification, accurately categorizing individuals into various credit score categories is paramount for financial institutions. This process directly influences financial opportunities, risk assessments, and loan approval decisions. By automating and enhancing the credit assessment process, a solution can lead to more prudent lending choices, reduced default risks, and improved financial management, benefiting both individuals and institutions.

Conversely, House Price Prediction plays a pivotal role in the real estate industry. Accurately estimating property values empowers real estate professionals, investors, buyers, and sellers to make well-informed decisions. It simplifies pricing strategies, streamlines real estate transactions, and maximizes returns on property investments.

To support these decisions effectively, I have diligently procured relevant datasets from reputable sources like Kaggle. These datasets encompass a wide array of features, including numerical and categorical variables, allowing us to harness the power of machine learning models for classification and regression tasks. The insights derived from these datasets are instrumental in enhancing decision-making processes, mitigating risks, and improving real estate and financial management.

In essence, my objective is to deliver tangible benefits to individuals, financial institutions, and real estate professionals through data-driven solutions. These solutions enhance decision-making consistency, precision, and efficiency, ultimately leading to improved financial management and more informed real estate transactions.

# References

Alam, S. *et al.* (2022) “Classification accuracy comparison between machine learning algorithms and a deep learning algorithm in predicting hand gestures,” in *2022 32nd Conference of Open Innovations Association (FRUCT)*. IEEE, pp. 22–29.

Dubey, A. and Rasool, A. (2019) “Data Mining based Handling Missing Data,” in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, pp. 483–489.

Houari, R. *et al.* (2014) “Handling missing data problems with sampling methods,” in *2014 International Conference on Advanced Networking Distributed Systems and Applications*. IEEE, pp. 99–104.

Jaiswal, J. K. and Samikannu, R. (2017) “Application of random forest algorithm on feature subset selection and classification and regression,” in *2017 World Congress on Computing and Communication Technologies (WCCCT)*. IEEE, pp. 65–68.

KUMARdatalab, H. (2023) “Housing price prediction.”

Mandala, S. K. (2023) “Credit Score Classification Dataset.”

Nazarov, F. M. and Yarmatov, S. (2023) “Optimization of prediction results based on ensemble methods of machine learning,” in *2023 International Russian Smart Industry Conference (SmartIndustryCon)*. IEEE, pp. 181–185.

Pathak, S., Mishra, I. and Swetapadma, A. (2018) “An assessment of decision tree based classification and regression algorithms,” in *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*. IEEE, pp. 92–95.

Safiya Parvin, A. and Saleena, B. (2020) “An ensemble classifier model to predict credit scoring - comparative analysis,” in *2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*. IEEE, pp. 27–30.

Sagi, S. L. N. *et al.* (2021) “Comparative analysis of data preparation methods employed in prediction of diabetes mellitus diagnosis using eICU-CRD dataset,” in *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. IEEE, pp. 1–6.

Torkjazi, M. and Raz, A. K. (2023) “Data-driven approach with machine learning to reduce subjectivity in multi-attribute decision making methods,” in *2023 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–8.