

**CAPSTONE PROJECT**

Assignment 4

**Analyzing and predicting rent prices in Canada**

*Professor: Samer Al-Obaidi*

*Submitted on 28th July 2024*

**Group: Ottawa**

**Team Members**

|  |  |
| --- | --- |
| **NAME** | **Student ID** |
| Tanya Nimesh | N01579781 |
| Yshika Shrikhande | N01545656 |
| Apeksha Hipparagi | N01579432 |
| Shrutika Desai | N01580471 |
| Gurleen Banga | N01579137 |

|  |  |  |
| --- | --- | --- |
| **Table of Content** | | |
| **S no.** | **Content** | **Page Number** |
| 1. | Introduction | 3 |
| 2. | Implementing the solution | 3 |
| 3. | Outcome Testing and reviewing | 7 |
| 4. | Optimization | 11 |
| 5. | References | 15 |

|  |  |  |
| --- | --- | --- |
| **Group contribution** | | |
| **S no.** | **Content** | **Stakeholder** |
| 4.1. | Implementation | Shrutika, Yshika, Tanya |
| 4.2. | Testing and reviewing | Tanya, Gurleen |
| 4.3. | Optimization | Apeksha |

**Introduction:**

The goal of this report is to develop an accurate system for analyzing and forecasting rent prices across Canada. Leveraging data from the Canadian Government website (Statistics Canada. (n.d.)), project aims to create a new IT architecture, as the existing system does not grant access to external parties. The implementation process involves several key steps, including data collection, data integration, and the application of machine learning techniques. This comprehensive approach ensures that the solution is well-integrated with the new IT infrastructure. The following sections detail the implementation process, including reflections and observations gathered throughout the project's development.

**4.1 Implementation of the Solution.**

Amid the vibrant and ever-changing landscape of Canada's real estate market, stakeholders sought a sophisticated tool to help them understand the intricate web of rent prices across various regions and housing types. The mission of this project was straightforward yet ambitious: delve into historical rent prices in regions with populations exceeding 10,000, predict future rent trends, and deliver actionable insights for real estate professionals, policymakers, and business owners alike. To tackle the pressing issues of the dynamics of market demand and supply, a robust solution is envisioned. The solution contains a relational database, a time series model, and an interactive dashboard. It equips stakeholders with practical, data-driven insights into regional rent comparison, market conditions, and future rent price predictions. The solution developed comprises several key components, each designed to provide a holistic and integrated approach to rent price analysis and forecasting. It is a Regression model along with an interactive dashboard on Power BI. As an assessment and visualization tool, this dashboard provides the stakeholders with a convenient way of interacting with different rental housing results. Such predictive analytics is of significant benefit to users because it can enable them to be prepared for changes in the market and make intelligent decisions.

Key Features of the Business Solution:

* Data Driven Insights: Utilize historical rent price information, helping in giving accurate and pertinent future predictions. Uses trends and patterns, among other factors, to give an accurate estimate of rent prices in the future.
* Regional Comparison: Highlight regions with rent prices significantly higher than the national average.
* User Friendly Interface: The layouts of both the sites were created with the real estate professionals, policymakers, and different business owners in mind. Easy to navigate with clear visuals and enhance user experience.
* Scalable and Flexible Solution: Designed to handle more of the changing needs and demands of the organization. Monitor the number of new rental housing units being produced. Flexible scale with new data inputs and analysis measures to better serve the user.

**Detailed Solution Design steps**

**Data Collection and Pre-processing:**

The historical rent price data is gathered (2020-2023) from The Canada Mortgage and Housing Corporation (CMHC) website (Statistics Canada. (n.d.))*.* For cleaning and pre-processing the dataset, PowerBI is used. Power BI is great because it can connect to various data sources like databases, spreadsheets, and cloud services, making it easy to bring in raw data for cleaning and transformation. Inside Power BI, there is a tool called Power Query that lets you clean, reshape, remove duplicates, split columns, and apply transformations using a visual interface. One of the best parts of using Power BI is that once you have set up, cleaning and preprocessing steps in Power BI, can be automated. This means your data will automatically refresh at regular intervals or whenever new data is added. This ensures that every time you analyze your data, you are working with the most up-to-date and cleanest information available.

**Model Development:**

Categorical Features: One-hot encoding is used to convert categorical variables into a binary format.

Numerical Features: Standard scaling is used to normalize numerical data.

Model Selection: Several different models were evaluated to determine which one worked best with the data. **The Gradient Boosting Regressor, Random Forest Regressor, CatBoost Regressor, and LightGBM Regressor** were tested. Each model underwent thorough training and testing, with the data split into 70% for training and 30% for testing to ensure that the models were evaluated on new data they had not seen before. Performance was measured using **Root Mean Squared Error (RMSE)** metrics, providing insights into the accuracy and dependability of the predictions. It seemed that the Random Forest Regressor could be among the best choices for further selection because of the trees which give accurate estimations. Its accuracy and assessment speed were favorable because one hundred estimators were employed. The seed number was selected to be a fixed random state of 42; this was because using the same seed number in the code produced the same sequence of random numbers as before, this was helpful in checking solutions against each other. For better understanding, visuals were made to allow people to easily decipher the variety and depth available. These visualizations pointed out trends and change anticipation, providing the general view for politicians and owners of existing enterprises.

The following are the benefits of using Random Forest Regressor:

* Accuracy: In the examined models, the Random Forest Regressor has the lowest value of RMSE therefore this model is least deviant from actual values and better determines patterns of data set than others.
* Robustness: From the above explanation, one can deduce that Random Forest is robust and can work effectively with almost any input data type and dimensions. Moreover, it minimizes overfitting and increases the theory’s generalization because it includes several decision trees.
* Interpretability: Random Forests also provide feature importance making it easier to identify which features have the most significant impact on the predictions even though the model is not as interpretable as the linear model.
* Scalability: Random Forests do not have problems with scalability, especially if they are designed to work on parallel computations as training is computationally intensive. However, Random Forests are slightly easier to use than some of the other complicated models like Gradient Boosting because they involve lesser parameter optimization.

**Steps for Client Company Personnel to Use the Developed Solution Successfully**

1. Access to Data and Data Integration

Updating the database often with the latest information on rent prices and pertinent indicators. In Power BI, setting up and keeping an eye on automated data, refreshing schedules to ensure the dashboard always shows the most recent data.

1. Using the Interactive Dashboard

* Impact Assessment: Using interactive filters to modify demographic characteristics and examining impacts based on selected criteria. Visualizing the effects of rising rent costs on various demographic groups.
* Regional Comparison: Navigating to the relevant dashboard section will get you access to the regional comparison tools. Assess regions to find locations where rent is noticeably more than the national average, then comparing and visualizing the results.

1. Predictive Model Utilization

Navigating to the area where future rent trends are provided by predictive models and analyzing the projected rent amounts for 2025, 2026, and 2027.

1. Generating Reports

To help with decision-making, using the dashboard's reporting tools to create custom reports based on needs or certain criteria.

1. Regular Training and Updates

Ensuring all users receive frequent training so they can keep informed about new features and industry best practices.

1. Monitoring and Maintenance

Ensuring the information displayed is accurate and reliable by doing regular data quality checks and monitoring the dashboard's and the underlying database's performance on a regular basis.

1. Utilizing Visualizations for Decision Making

Making use of visualizations to comprehend patterns and anticipated shifts, offering policymakers and company owners a road map. Making well-informed decisions on rent pricing strategies, policy development, and company planning by utilizing the dashboard's findings.

1. Feedback from Users

Providing a feedback system so that users may report problems or offer dashboard enhancement suggestions.

**4.2 Outcome testing and reviewing**

The Random Forest Regressor, Gradient Booster, Catboost Regressor, and LightGBM Regressor that are used to build models have been thoroughly tested. With the lowest root mean square error (RMSE) of all the studied models, the Random Forest Regressor can identify data patterns most effectively and generate the most accurate predictions. Among the models examined, the Random Forest Regressor results in accurate predictions with the lowest RMSE value of 89.258, as previously mentioned. This model testing, which is provided with insights, helps opt for the most effective model.

**RMSE vs** **MSE:** RMSE enables easy interpretation since the outputs are displayed in unit of the response variable. The purpose of choosing RMSE was that it penalizes larger errors more than smaller ones due to the squaring of the differences. This can be beneficial in scenarios where large errors are particularly undesirable. In this project, if RMSE is 100 dollars, it means that on average, the predicted rent prices are off by about 100 dollars. MSE, on the other hand, would be in squared units (dollars squared), making it harder to interpret.

Prediction of the Rent values for the years 2025, 2026, 2027 is done by the following 4 predictive models and following RMSE is found:

|  |  |
| --- | --- |
| Model | RMSE |
| Random Forest Regressor | 89.259 |
| Gradient Booster | 153.938 |
| CatBoost Regressor | 99.157 |
| LightGBM Regressor | 101.721 |

Taking all the values above into consideration, the RMSE is the lowest at **89.259** among all the models where the **Random Forest Regressor** was applied.

**Comparing the outcomes of different approaches:**

This ensured that the forecasts were exact and did not allow for variations apart from the level of complexity that was understood during the preprocessing phase and evaluation of several models. It provided at least a good structure, which was open for suggestion and further enhancement if better information was obtainable. Thus, this approach gave more detailed information because more complicated forms of relations within a dataset are considered in contrast to simpler forms like linear ones.

**Why Random Forest Regressor was chosen:**

Robustness: It is less prone to overfitting compared to boosting methods, especially when the number of trees is large.

Feature Importance: It provides a straightforward way to understand the importance of different features, which is valuable in understanding the drivers behind rent prices.

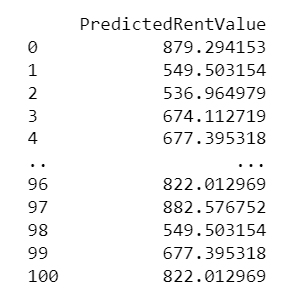
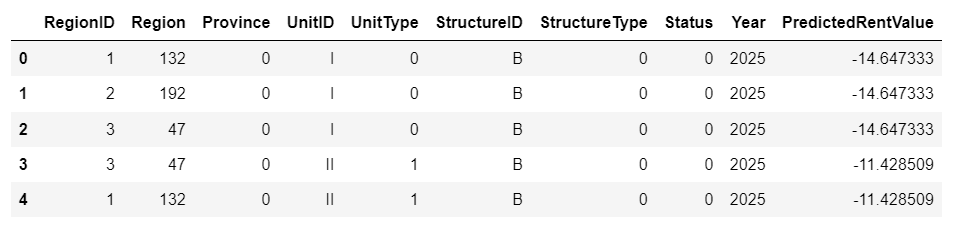
Interpretability: While not as interpretable as linear models, it offers a good balance, making it easier for stakeholders to understand the model's decisions.

Ease of Use: Random Forest Regressor is simple to set up and requires fewer hyperparameters to tune compared to boosting methods. This plays a crucial role in determining the future rent prices since the project’s aim is to provide actionable insights without getting bogged down in complex tunings.

Generalization Capability: Random Forest Regressor is more ensemble in its approach where the final decision is the average of many decision trees. This generalization is very important for the accurate future projections, a key objective of the project.

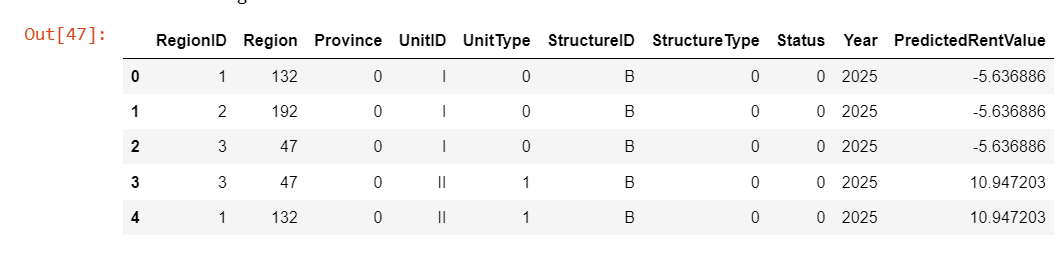
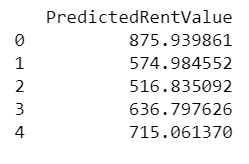
**Gradient Boosting Regressor:** It’s sensitivity to hyperparameter settings can lead to overfitting if not tuned correctly and understanding of its specific features is needed. It is less interpretable than Random Forest because of its sequential nature and the interactions between boost. It takes up more memory space than other algorithms while Random Forest being parallel is less computational, using less memory.

**BEFORE**  **AFTER**



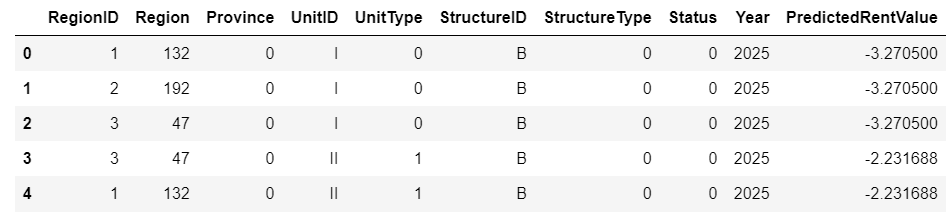
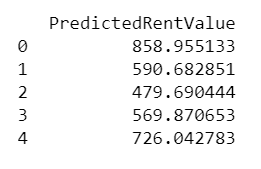
**Catboost Regressor:** Catboost is known for specifically handling categorical data and is less commonly used and may have less resources available. While it can be very effective and offer similar performance, it may require more tuning of its specific features.

**BEFORE**  **AFTER**

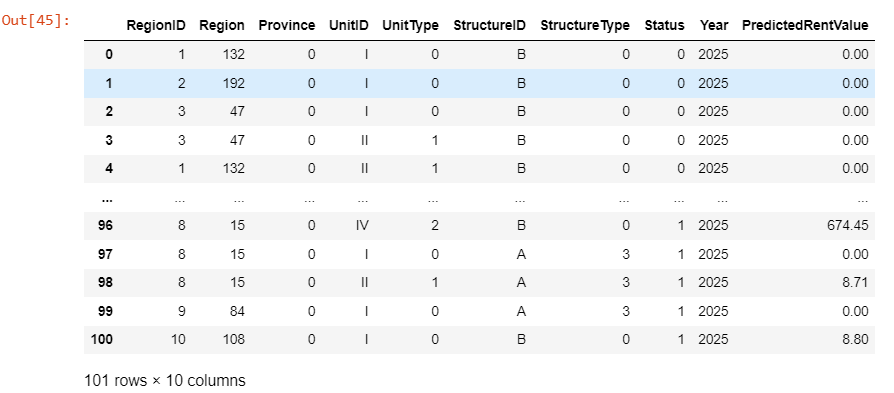
 

**LightGBM Regressor:** LightGBM is highly capable of handling large datasets, but compared to others, LightGBM incorporates new hyperparameters that require great attention to get the best results, and the whole process is very tedious. Additionally, LightGBM is a boosting method and involves building models sequentially, which can cause overfitting.

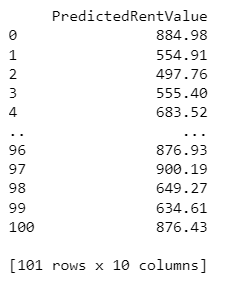
**BEFORE AFTER**

**Outputs for the code before the 70-30 percent split:**



**Output for the code after the 70-30 percent split:**



**Comparing results to the desired output:**

Desired Output: The target was to get close estimates of the rent prices and truly minimal error while choosing a model that is accurate, strong, and easy to work with.

Random Forest Regressor: This shows that Random Forest Regressor provided the smallest RMSE hence signifying high accuracy in the predictions. This makes it to be less sensitive to overfitting and thus extremely useful for giving future forecasts. Further, it is very convenient to train and adjust akin to several other models, and it involves the use of fewer hyperparameters. The model also helps to uncover the importance of features, and thus, identify the main factors influencing the rent level.

Comparison: Closely in line with the desired output by giving low forecasting error. It does not over fit easily to the training data; the training process is simple it also has few parameters which makes model management highly efficient. In addition, it provides information on feature importance for the analysis of specific attributes that influence rent prices; thus, the model can be considered accurate, robust, and easily navigable by the end-users.

**4.3 Optimization**

While the developed solution using the Random Forest Regressor allowed for the creation of reliable and accurate predictions for the rent price across Canada, there is always the potential for further refinement and enhancement. This section describes the directions and methods that could be applied to improve the performance, precision, and practicality of the solution.

**Enhancing Model Accuracy:**

1. Hyperparameter Tuning: Although the Random Forest Regressor was chosen because of its simplicity and efficiency, one can fine-tune the model’s parameters for better performance. Grid Search or Randomized Search with Cross-Validation allows choosing the best combination of parameters, for example, n\_estimators, random\_state, etc. This will aid in searching for the configuration that gives the lowest RMSE.
2. Feature Engineering: It was also found that adding more derived features to the dataset might have a positive effect on the model. For instance, feature selection can be incorporated in the pipeline using methods such as Recursive Feature Elimination (RFE) or Recursive Feature Elimination with Cross-Validation (RFECV). This approach ensures that only the most prominent features are used in the final model, potentially improving performance and interpretability.

**Data Quality and Integration Enhancements:**

1. Improving Data Refresh and Automation: Setting up weekly or monthly data refresh schedules guarantees that the model in Power BI is working with the latest data. Furthermore, all the steps, starting from data collection and ending with data preprocessing, can be automated, which would reduce the impact of human mistakes and maintain the data quality high.
2. Real-Time Prediction Capabilities: Adding a feature where a user enters specific situations (e. g. changes in interest rates or economic conditions) and receives an immediate prediction of the subsequent effects on the rent level will improve the tool’s usefulness in decision-making.

During the testing and review of the RMSE for the Random Forest Regression model, several errors were encountered. The table below summarizes the differences between the initial code and the final corrected code, along with the errors faced:

|  |  |  |
| --- | --- | --- |
| **Code** | **Initial** | **Final** |
| **Data Splitting** | X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42) | X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42) |
| **New Data Preparation** | unique\_combinations = df.drop(columns=['RentID', 'RentValue', 'Year']).drop\_duplicates() | unique\_combinations = df.drop(columns=['RentID', 'RentValue', 'Year']).drop\_duplicates() |
| **Added variability by using a trend factor based on historical growth** |  | historical\_growth\_rate = (df.groupby('Year')['RentValue'].mean().pct\_change().mean())  new\_data['PredictedRentValue'] = new\_data.apply(  lambda row: row['PredictedRentValue'] \* (1 + historical\_growth\_rate) \*\* (row['Year'] - 2024),  axis=1  ) |

**Key Differences and Issues:**

1. **Test Size:** The initial code used test data of 20%, while the final code was split into 70% training and 30% testing size. This change significantly affected the predicted values and found differences in the evaluation metric (RMSE) slightly.
2. **Model Evaluation Output:** During the initial analysis, it was found that the predicted rent values in the dashboard were the same for all future predicted years. To solve this issue, the above table of code was added which estimates the historical trends of the rent values and applies this on the projected rent values for the subsequent years in respect to this growth rate.

The pct.change() method calculates the percentage change between the current and prior element, providing the year-over-year growth rate in average rent values. Followed by this, it calculates the mean of these percentage changes, giving the average growth rate in rent values over the years.

This optimization step provides a more accurate prediction of future values of rent based on the previous ones. This is particularly helpful to the stakeholders that want to know about probable future market conditions.

**References**

* Data Source: Statistics Canada. (n.d.). Canada Mortgage and Housing Corporation, average rents for areas with a population of 10,000 and over. Retrieved July 18, 2024, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410013301>