Data Driven Decision Making

EXAM03: Data Science Group Assignment

Group:	The 3rd Bulgarian empire

Student names	Student numbers
1. Silvia Popova	1. 95653
2. Georgi Chitarliev	2. 95567
3. Chavdar Tsvetkov	3. 98008
4. Simeon Atanasov	4. 95616
5. Dimitar Petrov	5. 93425

UNIVERSITY OF APPLIED SCIENCES

WWW.HZ.NL

Iteration 2

BUSINESS UNDERSTANDING

Situation Understanding:

The HomeFirst Enchantment program was started by the High Council of Realms to help low-income families in locating affordable dwellings across several enchanted enclaves. Affordable homes, according to the council, are ones whose yearly housing expenses don't surpass 30% of a household's income. They require assistance in creating a classification model that will help them assess if the property is cheap to support this project.

Business Objectives:

By utilising a classification model, the High Council of Realms can effectively determine how affordable housing is. That will guarantee that families are guided towards appropriate dwellings that fulfil the affordability standards. Among the business goals are the following: maximising housing policy efficiency through data-driven recommendations; transparency in the housing market by defining clear rules to determine affordability; and improving decision-making for council officials and real estate consultants by classifying dwellings based on affordability.

Data Mining Goals:

Creating a basic rule-based classification model for determining a home's affordability is the main objective of data mining. This involves figuring out the main elements that affect affordability. Analyzing affordability trends across different enclaves to identify areas with a shortage of affordable housing is another objective. Through an optimization of the decision rules, the model can be continually improved to boost classification accuracy.

Success Criteria:

To ensure that most of the predictions match real affordability, a good model should be able to categorize the most affordable homes with an accuracy of at least 80%. For the model to efficiently identify inexpensive homes, precision and recall should both exceed 85%. To prevent classification mistakes from having a major influence on decision-making, the misclassification rate should stay below 20%. Furthermore, to ensure that the model continues to perform well overall in finding inexpensive properties while minimizing classification mistakes, the F1 Score should be at least 0.85.

DATA UNDERSTANDING – DATA EXPLORATION

The add-on dataset introduces 2 additional columns ('**StructuralIntegrity**'**,** '**ExteriorCharm**') to our existing dataset. They provide insight into the condition and appeal of the dwellings. While these factors are not directly related to finances, well-maintained homes may have higher trade values, while homes that are visually appealing but structurally weaker might be less affordable.

DATA PREPARATION – MERGING, VARIABLES, TARGET

After cleaning the Iteration 1 dataset, it should be exported and re-imported as ***set8_it1_cleaned.csv*** (exported in *Team 8: the 3rd Bulgarian empire - Iteration 1 Notebook*). The add-on dataset (***df_it2***) should be checked for duplicates, which can be easily removed. Then both datasets are ready for a left merge using the DI column as the key: ***df = df_it1.merge(df_it2, on='DI', how='left')***. We merge on the left so any outliers we have dropped will not be visible in the new dataset. Next, we define the key financial variables from the assignment (See *Team 8: the 3rd Bulgarian empire - Iteration 2 Notebook*). The variables are ***annual_income, monthly_income, interest_rate_year, interest_rate_month, loan_term,*** and ***total_payments.*** We also created separate columns for the '**Downpayment**'**,** '**LoanAmount**'**,** '**MonthlyPayment**'**,** and '**AffordableDwelling**' to be able to easily track those values.

DATA UNDERSTANDING

The newly merged and expanded dataset consists of 1 991 rows, with an average trade value of 166 555 and an average monthly loan payment of around 636. That shows that a lot of the dwellings (74.8%) are classified as affordable. However, there is a big variety in property values and loan amounts, with some dwellings reaching up to 337 000 in trade value and loan monthly payments going over 1 287 - that suggests affordability difference in different enclaves.

	Trade-Value	Living-Quarters	Parcel-Size	Structural-Integrity	Down-payment	Loan-Amount	Monthly-Payment	Affordable-Dwelling
mean	182,213	1,115	10,060	5.6	36,442	145,771	696	0.693
std	82,065	386	7,158	1.108	16,412	65,652	313	0.461
min	12,789	256	1,470	1	2,557	10,231	48.85	0
25%	129,900	879	7,410	5	25,980	103,920	496.13	0
50%	160,500	1,110	9,378	5	32,100	128,400	613	1
75%	213,220	1,329	11,546	6	42,644	170,576	814.36	1
max	755,000	4,339	164,660	9	151,000	604,000	2,883.59	1

Table 1: Mean, std and quartile summary

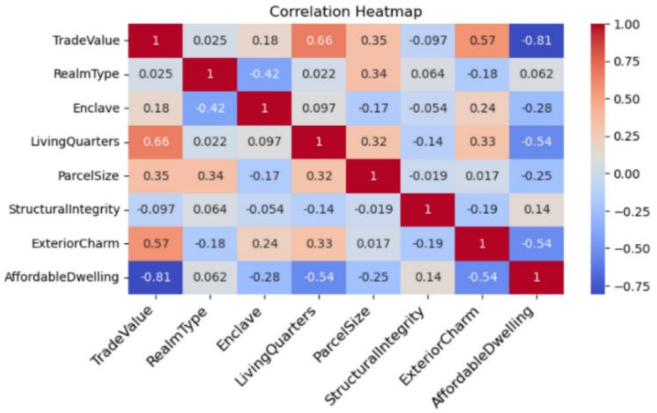


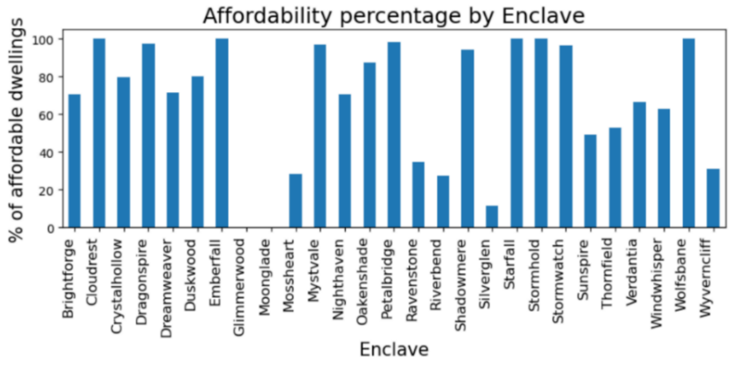
Figure 1: Correlation Heatmap of the main variables

Data Insights:

To highlight key relationships between variables we can use the correlation heatmap.

The primary metric for affordability is based on whether monthly loan payments exceeded 30% of a household's monthly income. That is the reason for the strongest negative correlation is between '**AffordableDwelling**' and '**TradeValue**' ***(-0.81)*** - meaning that as TradeValue increase, affordability significantly decreases.

Other insights include a moderate negative correlation *(*-0.54) between **'ExteriorCharm'** and '**AffordableDwelling**'. That suggests that visually nicer homes tend to be less affordable. Additionally, '**LivingQuarters**' (-0.54) also shows a negative relationship with '**AffordableDwelling**', which implies that bigger homes are generally less affordable. Interestingly, '**StructuralIntegrity**' has a weak positive correlation (0.14) with affordability - better-maintained homes may slightly contribute to affordability but are not a major factor. These insights show that '**Enclave**', '**LivingQuarters**', and '**ExteriorCharm**' are the strongest predictors of affordability.



On the connection between affordable dwellings and enclave, another key visualization is a bar chart showing the percentage connection. It appears that there is a significant variation of the number of affordable dwellings between different enclaves. The affordability is strongly influenced by '**TradeValue**', with enclaves that have lower median dwellings' prices having higher affordability rates.

Figure 2: Affordability percentage by enclave boxplot

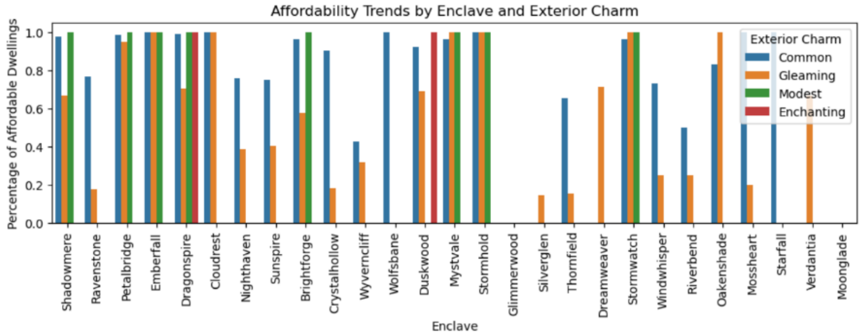


Figure 3: Affordability trends by enclave and exterior charm boxplot

Adding the '**ExteriorCharm**' as a feature we can see how affordability trends vary across enclaves and different levels of exterior charm. In many enclaves, homes with "**Common**" or "**Modest**" exterior charm tend to have the highest affordability rates, while homes with "**Gleaming**" or "**Enchanting**" charm are generally less affordable. Some enclaves, such as **'Ravenstone'** and 'Crystalhollow', show an evident affordability gap, where homes with higher charm levels are much less accessible, indicating that aesthetic appeal may drive up trade values and reduce affordability.

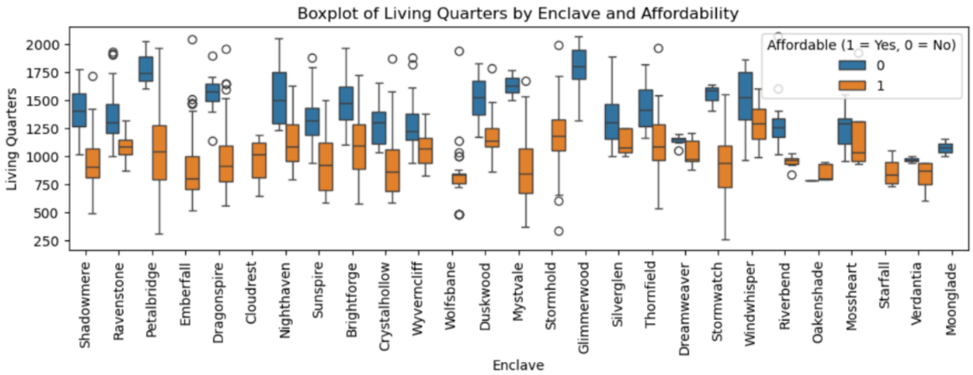


Figure 4: Affordability trends by enclave and living quarters boxplot

That is why we need to investigate '**LivingQuarters**'. In many enclaves, non-affordable homes tend to have significantly larger living spaces, while affordable homes generally have smaller median sizes. Some enclaves, such as **'Stormwatch'** and 'Shadowmere', show a clear separation between affordable and non-affordable homes, whereas others, like **'Mossheart'** and **'Moonglade'**, have a more balanced distribution, indicating that affordability in those areas may be influenced by factors beyond home size.

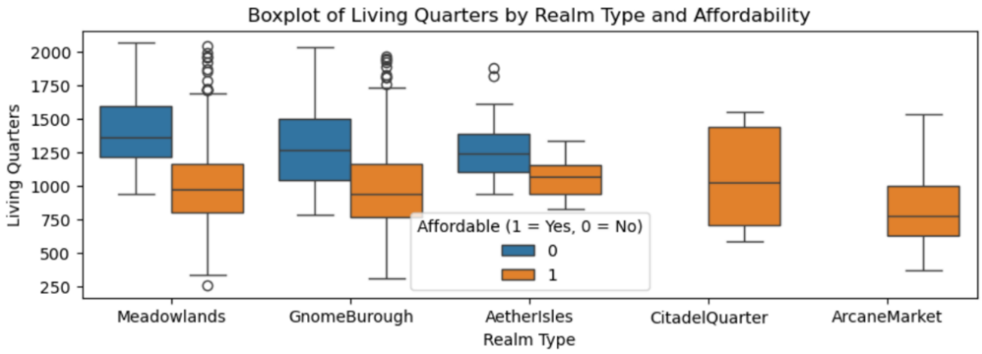


Figure 5: Affordability trends by real type and living quarters boxplot

Expanding on our findings, certain realms tend to have larger homes that are mostly unaffordable, while others show a greater mix of home sizes. In Meadowlands and **'GnomeBorough'**, larger homes are typically less affordable, while smaller homes in these realms have a higher likelihood of being classified as affordable. '**CitadelQuarter**'and '**ArcaneMarket**', on the other hand, consist entirely of affordable homes, suggesting that either these realms have smaller average home sizes or different economic conditions that make housing more accessible.

MODELLING

We will build a rule-based classification model because affordability follows clear, structured patterns based on enclave location, realm type, and property characteristics.

First, we must split the set into 70% training and 30% testing sets, separated into x and y respectively. For this, we define the target and features first: ***target = df3.pop('AffordableDwelling') & features = df3.***
And then we split: ***feature_train, feature_test, target_train, target_test = train_test_split(features, target, train_size=0.7, test_size=0.3, random_state=42)***

To ensure accurate predictions, we must define classification rules based on the insights gained from the Data Understanding phase. We can conclude that:

- The realm types '**CitadelQuarter**' and '**ArcaneMarket**' are always affordable.
- Specific enclaves are always affordable and some are never affordable.
- The modest exterior charm is always affordable.
- The enchanting exterior charm is always affordable on 2 enclaves, never otherwise.
- The common/gleaming exterior charm is always affordable on some enclaves and partially on others (done with random), and rarely not affordable.
- Mostly when the Living Quarters are more than 1500 the property is likely not affordable.

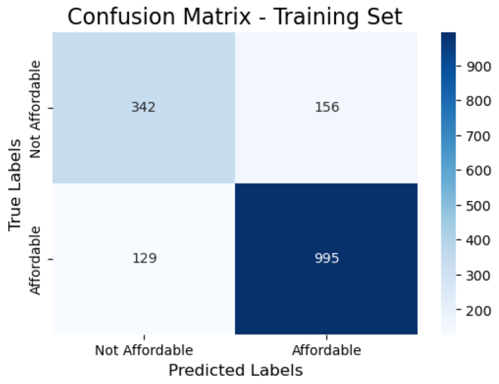
In order to provide a more accurate representation of borderline cases, we introduce randomness into our model to account for minor affordability changes that are impossible to capture by strict rules alone. Thus the results will always vary slightly.

The final step is we have created the classification predictive model. We are using a lot of if statements, nested if statements and random. Random is used to simulate as much as possible the nuance of affordability without making a more complex algorithm. It does add slight bias (rounding the numbers), but it is negligible. *(See Team 8: the 3rd Bulgarian empire - Iteration 2 Notebook for full function)*.

After the predictions we are checking the accuracy with ***accuracy_score*** from ***sklearn.metrics*** to see what the efficiency of our model is. The results will always vary slightly (due to random), but on average the **accuracy should be around 82%** (more explained in the next chapter: Evaluation)**.**

EVALUATION

After the model is being created and trained it comes time to how efficient it is. We create a confusion matrix, which looks like the following:



The metrics we are using to determine if our model is efficient are:

- **Accuracy**: Measures the proportion of correctly classified dwellings out of all predictions. We use this metric to evaluate the overall performance of the model.
- **Precision**: Indicates how many of the predicted affordable dwellings are affordable. This is important to minimize false positives and avoid incorrectly labeling unaffordable homes as affordable, which could mislead low-income families into considering properties beyond their financial capabilities, causing financial issues.
- **Recall**: Measures how many of the truly affordable dwellings are correctly identified by the model. We use this to ensure that we do not overlook affordable homes, which could prevent families from accessing suitable housing options.
- **F1 score**: Represents the balance between precision and recall - a single metric to assess both false positives and false negatives. This is useful for evaluating the model when both types of errors carry significant consequences, as is the case with our project (as explained in the previous points).
- **Misclassification rate**: Shows the percentage of incorrect predictions made by the model. We track this to minimize errors.

Based on these metrics, our model successfully meets the defined success criteria, demonstrating a very good performance in affordability classification. Here are our results:

- Accuracy of 0.82 - good overall performance that the model correctly classifies the majority of dwellings
- Precision of 0.86 - the model predicts dwelling as affordable 86% of the time
- Recall of 0.89 - the model is identifying 89% of the truly affordable dwellings
- F1 Score of 0.87 - there is a good balance between precision and recall
- Misclassification rate of 0.18 - the model makes an error 18% of the time, meaning it still maintains a high level of accuracy

To improve the model in the future we may need to dive deeper into some aspects of our model. The first aspect is better understanding of the misclassification of our features so we can reduce the misclassification rate to a minimum. Secondly, we may need to focus better on the data preparation stage, where we can change the values of the outliers with the median values. The affordability classification model will be further improved, and performance will be maximized with these improvements.

PERSONAL CONTRIBUTION

	Student name	Contribution
1	Georgi Chitarliev	Modeling • Create a benchmark classification model using simple rules. • Improve the benchmark model by adding more rules, based on the findings of the data understanding.
2	Silvia Popova	Data Understanding • Include your data exploration (e.g., summary statistics, visualizations). • Summarize key trends in affordability patterns (e.g., percentage of affordable dwellings by enclave). Overall document structure check
3	Chavdar Tsvetkov	Business Understanding • Describe the current situation and define the problem and business objectives. • Define and justify the data mining goal(s) to be able to determine the success criteria for this iteration.
4	Simeon Atanasov	Evaluation • Evaluate the model using a confusion matrix and appropriate classification metrics – accuracy, precision, recall, F1 score, misclassification. • Discuss how well the benchmark model performs and identify areas for improvement.
5	Dimitar Petrov	Data Preparation • Merge, clean and prepare the dataset • Create calculation variables based on assignment and a binary target variable (AffordableDwelling) correctly following the formula stated in the case.