**5) MODEL SELECTION, TRAINING, AND TESTING**

This chapter details the model selection, training, and testing processes. It also contains empirical results and analysis.

**5.1) Price Prediction Model**

Model Selection

The baseline model provided in the project brief used a hedonic OLS regression, use of this model is continued to provide a baseline for more advanced modelling methods. The selection of advanced modelling methods was informed by the project requirements and findings from the literature review. Interpretability was seen as an important factor in the model selection process given the requirement to output the predictive importance of variables. Given ANNs 'black-box' reputation and inability to produce variable importance, it was decided that it would be more appropriate to use ensemble methods in property price predictions.

Two ensemble methods were discussed in the literature review, Random Forest with its use of bagging, and XGBoost and its use of boosting. Both models are trained on the dataset, and their results compared to one another as well as the baseline hedonic model. The same training dataset is used by all models.

Training, Testing, and Optimisation

Prior to model training the pre-processed unstructured data was split into two subsets. These subsets contained the unstructured variables for the most expensive and cheapest 25% of ads. The extracted nouns, adjectives and lemmatised bigrams for these subsets were then merged with the structured data to create the final dataset for the models, the count and average length of each word-type were also included. The final dataset contained 2,458 observations and 122 variables. 3 categorical variables, 13 numerical variables, and 106 binary dummy variables that indicate the presence of a word in the property's description or features. The dummy variables consist of 69 nouns, 15 adjectives and 22 lemmatised bigrams.

The dataset was split into training and a testing sets in order to evaluate the accuracy of the models. 80% of the data was used to train the models with the other 20% being used to test the accuracy of the models. Three metrics were used to assess model accuracy Median Absolute Percentage Error (MdAPE), Mean Absolute Error (MAE), and R-Squared. MdAPE and MAE will give us an indication as to how precise the model is, giving us a percentage and real number value. R-Squared indicates the variance in price that is explained by the independent variables and will give us an insight into the importance of different variables and the goodness of fit of the models.

The two ensemble methods, XGBoost and Random Forest, ingest several 'hyperparameters' in addition to the training data to tune their performance and correct overfitting. There are a number of ways to tune these hyperparameters, these include brute force methods such as 'GridSearch' or Bayesian optimisation methods using 'HyperOpt'. Bayesian methods have been shown to be faster and more accurate tuning models in practice and as such 'HyperOpt' was used to tune both ensemble methods.

Results

Two objectives were kept in mind when creating the price prediction models. The first was to provide the most accurate price prediction. The second objective was to assess the predictive importance of variables, specifically variables derived from the unstructured data. To accomplish both these objectives, the final dataset was incrementally fed to each of the three models. The incremental training process is outlined with the help of the XGBoost predictive results table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MdAPE** | **MAE** | **R-Squared** |
| **Structured (Base)** | 9.75 | 53415.80 | 0.823 |
| **Base + Nouns** | 11.52 | 61009.86 | 0.772 |
| **Base + Adjectives** | 11.11 | 64287.03 | 0.741 |
| **Base + Lemmas** | 11.07 | 56941.40 | 0.805 |
| **Base + Nouns + Adjectives** | 11.88 | 63700.74 | 0.771 |
| **Base + Nouns + Lemmas** | **8.69** | 47365.43 | **0.867** |
| **Base + Adjectives + Lemmas** | 9.26 | 51362.08 | 0.745 |
| **Base + All Word-Types** | 8.85 | **47202.34** | 0.864 |
| **Base + Counts + Lengths** | 9.45 | 51336.54 | 0.834 |
| **Base + All + Counts** | 9.76 | 51958.86 | 0.841 |

Firstly, structured data was used to train the model, unstructured variables are then incrementally added. This allows us to assess the predictive importance of nouns as compared to adjectives and so on.

Chart, scatter chart

Description automatically generatedXGBoost performs the best overall providing the most accurate predictions and outperforming the OLS and Random Forest models in every training iteration. The accuracy results for all three models can be compared in APPENDIX X. The best XGBoost model is highlighted above and consists of structured data, nouns, and lemmas. In the plot below we can see the predicted price as compared to the actual price for this model.

We can see that in the region from €200,000 to €400,000 that there is reasonably good fit with many datapoints centred around the 45-degree line. As the price rises however there is more varied results, and this can be seen when assessing the predictive accuracy by price range as seen in APPENDIX X. We can see that as the price rises the number of training rows starts to decrease meaning the model can't predict as accurately. Overall, this model provides reasonable accuracy and would be expected to improve if given a larger dataset to train on.

Importance of Variables

Table

Description automatically generatedIn completing the second objective, assessing the variable importance, we look firstly at the importance of different word-types. We can see from the comparable results in APPENDIX X, that the word-types add and sometimes detract from the model in varying degrees. Of the three word-types adjectives add the least to the model in isolation, improving the MAE and R-Squared of the OLS model only slightly and negatively impacting the XGBoost and Random Forest models across all metrics. Lemmas improve the OLS and Random Forest models across all metrics, but nouns are slightly better at improving model performance. Interestingly adding any word-type in isolation to XGBoost negatively impacts performance as compared to the base dataset with only structured variables. Nouns and Lemmas are the best combination and produce the best overall model. Including all word-types can also improve performance and produced the best Random Forest model and resulted in the lowest MAE in all three modelling algorithms.

After assessing the impact of different word-types a more granular approach is taken to assess the most important variables overall. The variable importance for the best model (XGBoost with Nouns + Lemmas) is shown in APPENDIX X. We can see that the structured data is more impactful than any of the variables derived from the unstructured descriptions. Unsurprisingly the geographical variables 'latitude' and 'longitude' are the most important predictors of property value in this model, echoing the "Location x 3" trope in real estate. Among the unstructured variables nouns have greater representation than lemmatised bigrams in the most important features supporting the table results in displaying their greater predictive power. \*Prefixes are added to unstructured variables in graphing to indicate word-type

By plotting the decision tree behind the XGBoost model in APPENDIX X, we can see the decision path of the model and how it derives a final predicted price for a property and the impact variables play in the prediction process. An example is highlighted by the black box, here we can see that where 'N\_wardrobes' < 0.5 (i.e., not present in the ad description) the final leaf node values are lower than those where 'N\_wardrobes' is present. This gives us extra insight into 'N\_wardrobes' relationship with price, its inclusion in the description can be important in prediction.

**5.2) Location Prediction Model (Classification)**

Model Selection

too little data too many classes

Multi-nomial Logistic Regression / Random Forest

(N\_clusters vs. Accuracy) how to add different NLP to this???.

Multi-output ANN to predict Lat and Long make this a regression problem

The price prediction model is an exercise in regression, predicting a number on a continuous scale. The location prediction model is an exercise in classification, predicting a nominal category or class associated with an observation, in this case the class is the property area. There are different requirements and constraints to both prediction exercises. In this classification task, the biggest constraint is the size of the dataset as compared with the number of classes. There are 2,458 observations in the final dataset, and 144 distinct areas. This large number of areas poses a problem, there are many areas with few observations. As we saw in the regression exercise the accuracy decreased when price rose as there were less training observations available. 45% of the areas have less than 10 observations, meaning the models will have too few training observations to produce an accurate output for many areas. To further illustrate this, three models were trained with the current dataset, a multinomial logistic regression model, an XGBoost Classifier, and a Random Forest Classifier, their respective accuracy scores were: 4.2%, 6.43%, and 12.6%.

To circumvent this problem two alternative approaches were taken. The first approach clusters the observations using latitude and longitude to create aggregated areas and reduce the number of classes. K-means clustering will be used to cluster the data and multinomial logistic regression, XGBoost classifier, and Random Forest classifier models will be used to predict the area cluster. The second approach changes this problem from a classification task to a multi-output regression task and attempts to predict both the latitude and longitude of properties. A Random Forest regressor will be used to perform the regression approach as it inherently supports multiple outputs.

Training, Testing, and Optimisation

In the price prediction model, the dataset was derived by extracting the nouns, adjectives, and lemmatised words from the top and bottom 25% of properties in terms of price. The location prediction dataset consists of the structured variables, nouns, adjectives, and lemmatised words/bigrams that occur in 30-60% of all observations. This is done as it would be extremely time consuming to create different bags of words for each area in the classes and would be difficult to apply in practice to a multi-output regression model for latitude and longitude.

The regression model will make use of the same metrics as used for the price prediction model, MdAPE, MAE, and R-Squared. The clustered approach will use the classification accuracy score as its measure of accuracy. The data will be split into training (80%) and testing (20%) data in the same manner as before.

All ensemble methods, XGBoost and Random Forest classifiers and regressors will be tuned using 'HyperOpt' as in the price prediction model.

Predictive Results (Cluster Approach)

Clustered aggregates of areas were created with number of clusters ranging from two to twelve. The optimum number of clusters using KneeLocator was 4, however more granularity would be required of a commercially viable product, and 12 was decided to be the maximum number of clusters allowed. Plots of the clusters can be seen in APPENDIX X. Different areas of Dublin can be made out in the plots with higher numbers of clusters, Dun Laoghaire and the coastal south-east, inner-city Dublin, Howth in the north-east, etc.

The accuracy scores of all three clustered models are shown in the table below. Across all three models the predictive accuracy falls as the number of clusters rises. This is a result of training observations being spread more thinly across a larger number of classes in 12\_clusters as compared with 2\_clusters for example.

|  |  |  |  |
| --- | --- | --- | --- |
| n\_clusters | Multinomial Logistic Regression | XGBoost | Random Forest |
| 2\_clusters | 72.07% | 79.06% | 75.15% |
| 3\_clusters | 58.26% | 69.40% | 58.93% |
| 4\_clusters | 57.70% | 66.74% | 54.62% |
| 5\_clusters | 49.50% | 63.04% | 50.72% |
| 6\_clusters | 48.51% | 62.22% | 51.33% |
| 7\_clusters | 46.20% | 59.55% | 46.41% |
| 8\_clusters | 42.94% | 57.29% | 46.00% |
| 9\_clusters | 37.04% | 49.49% | 40.86% |
| 10\_clusters | 42.10% | 51.13% | 41.27% |
| 11\_clusters | 39.00% | 50.51% | 38.81% |
| 12\_clusters | 38.12% | 54.00% | 37.58% |

Again, we see that XGBoost is the best performing model across all iterations, with Multinomial Logistic Regression and Random Forest performing similarly. When deciding the optimal model to use users of this approach face a trade-off decision between granularity and accuracy. The model that will be used to interpret the importance of variables is XGBoost with 6 clusters, this model offers reasonable granularity and is reasonably accurate.

Importance of Variables (Cluster Approach)

Table

Description automatically generated

Like the price prediction model, we can see that the structured variables provide more predictive power than the variables derived from descriptions. This feature importance is almost the inverse of that of the pricing model, where locational variables (latitude and longitude) were the most important features in predicting price, here price is the most important variable in predicting location. Some other interesting variables in the most important features are 'L\_city\_centre' rather intuitively it would indicate a property close to the city centre, 'L\_luas' would indicate proximity to the Luas, Dublin's tramline system. 'N\_view' also features, this could indicate a coastal property with 'sea views' perhaps. As compared with the pricing model, the unstructured variables are not as overshadowed by the importance of structured data in predicting location. This could suggest that descriptions are more powerful for predicting the location of a property than the property price.

Predictive Results (Regression Approach)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MdAPE** | **MAE** | **R-Squared** |
| **Latitude** | 0.04 | 0.0311 | 0.35872938 |
| **Longitude** | 0.61 | 0.0517 | 0.43027817 |

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedThe results for the multioutput Random Forest model in the table above look promising on the surface, with low MdAPE and MAE suggesting small error. However, when we put this in the context of a low variance target variable like latitude latitude these are large errors that could place the property kilometres away from its actual location. Visualising the output with scatterplots offers a true reflection on the accuracy of the model.

From the above plots we can see that the model is a poor fit for both outputs as datapoints deviate from the line of best fit. This is supported by the R-squared values for both latitude and longitude.

Graphical user interface, table

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generatedImportance of Variables (Regression Approach)

We can see that both outputs have differing variables of importance. Price features prominently in both as we might expect given its relationship to locational variables shown in the previous models. Again, we see that L\_luas and L\_city centre are important predictors of both latitude and longitude. The variables derived from the unstructured descriptions are of greater importance in this model similar to the cluster approach. This may further support the claim that descriptions are more useful in location prediction than price prediction.

**OLS Price Prediction Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MdAPE** | **MAE** | **R-Squared** |
| **Structured (Base)** | 17.07 | 86666.57 | 0.578 |
| **Base + Nouns** | 14.82 | 77208.55 | 0.674 |
| **Base + Adjectives** | 17.66 | 86526.70 | 0.586 |
| **Base + Lemmas** | 14.68 | 78407.18 | 0.643 |
| **Base + Nouns + Adjectives** | 14.76 | 77150.28 | 0.671 |
| **Base + Nouns + Lemmas** | 14.47 | 72638.56 | **0.713** |
| **Base + Adjectives + Lemmas** | 15.22 | 78854.16 | 0.646 |
| **Base + All Word-Types** | 14.42 | **72415.30** | 0.708 |
| **Base + Counts + Lengths** | 16.61 | 83522.98 | 0.604 |
| **Base + All Variables** | **14.22** | 72541.53 | 0.710 |

**XGBoost Price Prediction Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MdAPE** | **MAE** | **R-Squared** |
| **Structured (Base)** | 9.75 | 53415.80 | 0.823 |
| **Base + Nouns** | 11.52 | 61009.86 | 0.772 |
| **Base + Adjectives** | 11.11 | 64287.03 | 0.741 |
| **Base + Lemmas** | 11.07 | 56941.40 | 0.805 |
| **Base + Nouns + Adjectives** | 11.88 | 63700.74 | 0.771 |
| **Base + Nouns + Lemmas** | **8.69** | 47365.43 | **0.867** |
| **Base + Adjectives + Lemmas** | 9.26 | 51362.08 | 0.745 |
| **Base + All Word-Types** | 8.85 | **47202.34** | 0.864 |
| **Base + Counts + Lengths** | 9.45 | 51336.54 | 0.834 |
| **Base + All Variables** | 9.76 | 51958.86 | 0.841 |

**Random Forest Price Prediction Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MdAPE** | **MAE** | **R-Squared** |
| **Structured (Base)** | 14.18 | 69,051.50 | 0.738 |
| **Base + Nouns** | 13.96 | 68901.31 | 0.741 |
| **Base + Adjectives** | 14.32 | 69166.95 | 0.736 |
| **Base + Lemmas** | 14.02 | 68880.99 | 0.739 |
| **Base + Nouns + Adjectives** | 13.89 | 68778.86 | 0.743 |
| **Base + Nouns + Lemmas** | 14.05 | 68718.59 | 0.743 |
| **Base + Adjectives + Lemmas** | 14.12 | 68985.94 | 0.739 |
| **Base + All Word-Types** | 14.31 | **68630.96** | **0.745** |
| **Base + Counts + Lengths\*** | 13.89 | 69361.15 | 0.735 |
| **Base + All Variables** | **13.84** | 69226.73 | 0.739 |

\* Counts + Lengths: Counts are the number of each word-type in a description (e.g., count of nouns), Lengths are the average length of word-types in description (e.g., average length of adjectives).

**Accuracy vs. Price Range**

|  |  |  |  |
| --- | --- | --- | --- |
| **Price Range** | **MdAPE** | **# Testing Rows** | **# Training Rows** |
| **(50000, 100000]** | 215.58 | 1 | 0 |
| **(100000, 150000]** | 37.29 | 1 | 11 |
| **(150000, 200000]** | 11.03 | 30 | 104 |
| **(200000, 250000]** | 11.93 | 70 | 236 |
| **(250000, 300000]** | 10.12 | 74 | 306 |
| **(300000, 350000]** | 8.96 | 57 | 273 |
| **(350000, 400000]** | 8.46 | 61 | 234 |
| **(400000, 450000]** | 9.52 | 34 | 152 |
| **(450000, 500000]** | 8.5 | 38 | 159 |
| **(500000, 600000]** | 10.9 | 51 | 195 |
| **(600000, 700000]** | 7.81 | 25 | 87 |
| **(700000, 800000]** | 10.88 | 22 | 102 |
| **(800000, 1000000]** | 15.39 | 23 | 88 |

**Diagram

Description automatically generatedXGBoost Price Prediction Decision Tree**

Diagram, calendar

Description automatically generated**Cluster Plots**