Discovering Patterns in the Russian Housing Market for Analysis and Prediction

3804ICT Assignment Part I | Project Proposal | Trimester 2, 2019

Joshua Russell s5057545
Joshua Mitchell s5055278
Hayden Flatley s5088623

{joshua.russell2, joshua.mitchell4, hayden.flatley}@griffithuni.edu.au

1. Aim

The aim of this project is to (1) predict Russian house prices using information about the property and the surrounding area, (2) find relationships and associations between housing market attributes and properties themselves, and (3) develop a classifier for predicting the sub-areas of properties.

Predicting house prices would not only allow home buyers and sellers to more objectively evaluate the value of properties, but it would also give real estate businesses and banks additional knowledge to use for assisting customers. Furthermore, investigating what house market features are related or associated in some way will allow us to discover intrinsic patterns in the market that may prove useful for improving the predictive accuracy of our models in (1), or in understanding characteristics behind properties and the types of properties found within certain areas. Lastly, developing a model for classifying property attributes into sub-areas of Russia will provide home buyers with a way of finding their desired property area. Such a model could be used in an application that allows home buyers to specify their ideal property and neighbourhood features as input and receive recommended sub-areas to rent or buy in as output.

2. Data

We will use a Russian housing market dataset provided by Sberbank for this project. Sberbank is a Russian banking and financial services company that uses this information for helping their customers plan budgets when developing, renting, and buying properties. The primary dataset consists of attributes describing property features and information about the surrounding neighbourhood. These attributes consist of both categorial and numerical valued information. In total, this dataset has 292 attributes and 38,133 samples. Alongside the primary property dataset, Sberbank also provided a secondary dataset consisting of macroeconomic indicators of Russia. At the time of writing, we do not plan to use this secondary dataset in the final project investigation. However, it may be used to increase the predictive accuracy of regression models, or in clustering and association rule mining.

The dataset was made available online at Kaggle [1].

3. Algorithms and Techniques

Numerous data mining methods (regression, forecasting, association rule mining, clustering and classification) will be used to accomplish the desired outcomes discussed within *1. Aim*. For clarity, we will discuss each method under the heading of its corresponding desired outcome. Furthermore, for each method, the details of only one algorithm/technique will be introduced. However, additional algorithms/techniques may be used in the final project report and investigation.

3.1. Predicting Property Prices

Regression and forecasting methods will be investigated to predict the price of Russian properties. A *multilayer perceptron* (MLP) will be used for the task of regression. The MLP is a type of artificial neural network which learns a mathematical function $f(x, \theta)$ that, in the task of regression, maps a certain input (e.g. house property attributes) to a continuous valued output (e.g. house price). The MLP will be trained to learn this function through supervised learning. Where n training examples of inputs x and corresponding true outputs y are used to update the network's parameters θ by backpropagating the network's error to minimise a loss function such as mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \theta))^2$$

Exponential smoothing will be used for forecasting property prices. Broadly, forecasting techniques model the patterns found in previously observed values to make predictions about future values. Exponential smoothing takes this general idea one step further by applying an exponentially decreasing weight on each of the previously observed or forecasted values, so that its prediction is more responsive to recent trends in the data. Formally, for a given sequence of actual values $\{x_t\}$, the forecasted value s_t at a given time t is given by:

$$s_0 = x_0$$

when $t = 0$, and

$$s_t = \alpha x_t + (1 - \alpha)s_{t-1}$$

when $t > 0$ and where $\alpha \in (0, 1)$ is the smoothing factor.

Other forecasting techniques that may also be experimented with are moving average, weighted moving average, and seasonality analysis.

3.2. Discovering Relationships and Associations

To find relationships and associations among housing market attributes and the property data samples, clustering and association rule mining methods will be implemented. A popular partitioning algorithm, k-means, will be used for clustering. k-means iteratively groups data points into k clusters, minimising the sum of squared distances E between the cluster centroids c_i and the points $p \in C_i$ assigned to a particular cluster i:

$$E = \sum_{i=1}^{k}\sum_{p \in C_i}(p - c_i)^2$$

The algorithm achieves this minimization, given a specified k, by starting with an initial random k means (points). Each data point is then (i) assigned to the nearest k-mean, and (ii) a new mean point is calculated for each of the k clusters using the assigned data points. Steps (i) and (ii) are repeated until convergence (i.e. the data points assigned to the k clusters remain the same) or until a specified number of iterations is reached.

There are some weaknesses of the k-means algorithm. Such as the need to specify k and the fact that the algorithm is sensitive to noise and outliers. As a result, DBSCAN, a density-based clustering algorithm, may also be explored in the final project investigation.

Association rule mining involves two fundamental steps, frequent itemset generation and rule generation. We will use the Frequent Pattern-Growth (FP-Growth) algorithm to find all frequent itemsets within the dataset. FP-Growth uses a tree structure (FP-tree) to efficiently mine frequent patterns using a depth-first search approach. Avoiding explicit candidate itemset generation and requiring only one scan of the database for itemset support counting. Once the frequent itemsets have been found, association rules will be generated through a binary partitioning of the frequent itemsets. Rule evaluation metrics such as support, confidence and lift will then be used to determine which association rules are strong and/or interesting.

3.3. Classifying Property Sub-Areas

The task of classifying Russian sub-areas based upon property attributes will be achieved using classification methods. We will use a probabilistic classifier based on Bayes' theorem, known as the Naïve Bayes Classifier, for this task of classification. The classifier works by taking a feature vector $X = (x_1, \dots, x_n)$ and estimating the conditional probability of observing a class value C_k given the feature vector X:

$$P(C_k|x_1, \dots, x_n)$$

This conditional probability is estimated for each of the K possible class values, and the class with the maximum probability is chosen as the classifier's prediction. A key aspect of the Naïve Bayes Classifier is that it assumes that the features $x_1, \dots, x_n$ are independent, given the class variable, which makes calculating the estimated probability feasible in practice. However, if this assumption is violated then the performance of the classifier will suffer. Since our data contains both categorical and numerical attributes, we will need to use bin discretization or probability density estimation for numerical attributes in our implementation. In the final project investigation, classification algorithms such as k-nearest neighbours, the MLP and the decision tree may also be investigated.

4. Evaluation Measures

All algorithms/techniques implemented will be evaluated in terms of time complexity and correctness against library implementations. Furthermore, for the regression and the forecasting methods, *mean absolute error* and *mean squared error* will be used as evaluation measures. These measures quantify the difference between two continuous variables, which in our investigation will be the predicted house price and the true house price. We will use *support*, *confidence*, and *lift* to evaluate the interestingness of association rules. These metrics combine the occurrence frequency of itemsets and sub-itemsets in different ways to compute numerical representations of the usefulness, certainty and correlation of association rules, respectively. Clustering algorithms will be evaluated *intrinsically* through qualitative analysis. These intrinsic evaluations will involve examining common features in clusters to find interesting similarities and differences between houses and areas. *Extrinsic* evaluation measures may also be used. Such measures will evaluate the change in predictive accuracy of regression and forecasting methods when these methods are trained with data from within a particular cluster (in comparison to being trained with the entire dataset). Finally, the classification methods will be evaluated using common measures such as *accuracy*, and from the elements of the confusion matrix, *precision* and *recall*. Each measure provides an alternative approach for evaluating the predictive performance of the classifier on the test set.

5. References

[1] Sberbank. (2017, April). Sberbank Russian Housing Market. Retrieved August 20, 2019 from https://www.kaggle.com/c/sberbank-russian-housing-market