2023

Report - Lab 3

SOLVING CLASSIFICATION AND CLUSTERING PROBLEMS IN BI

TWAN GERJO VELDHUIS, H22TWAVE@DU.SE

Task 1: Descriptive Analysis, Unsupervised Learning – IKEA Case

In task 1, the goal was to find potential locations for IKEA department stores in Sweden using the k-means clustering method to categorize municipalities. The first step was to load the data file that contained all the relevant information. Next, we filtered out any categorical or text-based columns that were not necessary for the analysis.

```
# 1. Clean the dataset
file = 'C:/Users/tveldhuis/OneDrive - DigitalWorkPlaces/Documents/School/Business Intelligence/Lab
df_raw = pd.read_csv(file, sep='\t') # Replace with your actual file path
df = df_raw.select_dtypes(exclude=['object']) # Exclude categorical/textual data
df = df.drop(["Kommun_code", "Year", "Infrast", "Border"], axis=1) # Exclude categorical data
```

Figure 1 - import data and partially clean it

To streamline the analysis in task 1, irrelevant columns, including categorical and text-based data, were removed from the data file. Following this, a dimensionality reduction technique was applied to 95% of the data, resulting in a reduction in the number of columns from 7 to 3 without any loss of rows.

To determine the optimal number of clusters for the k-means prediction, the Elbow method was employed. This involved plotting the sum of squared distances between data points and their assigned cluster centers against the number of clusters, and then identifying the "elbow point" where the curve begins to level off. The number of clusters corresponding to the elbow point was chosen as the optimal value for the k-means prediction.

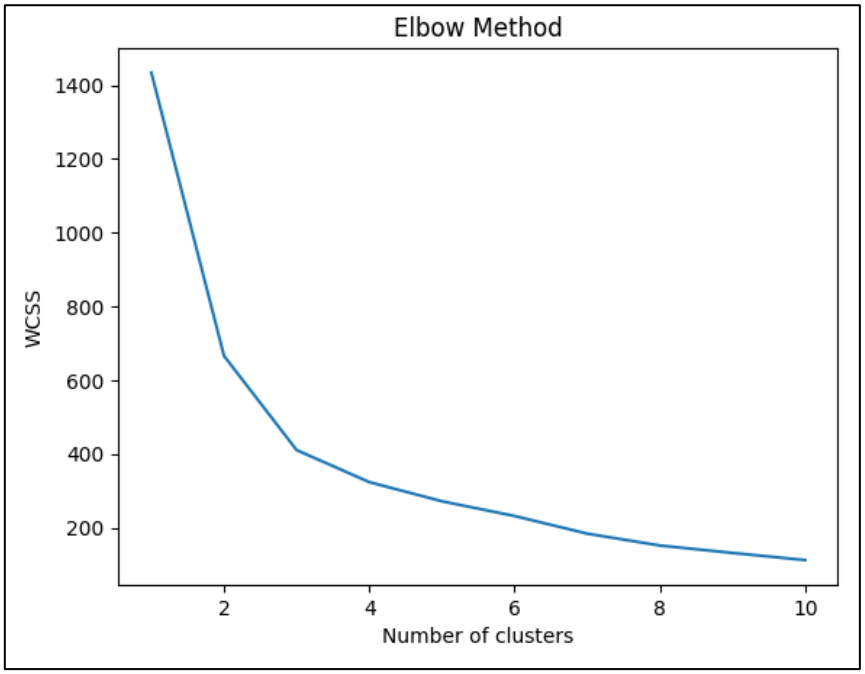

Figure 2 - elbow plot

This elbow method execution resulted in a number of 3 clusters for the upcoming analysis. This was then used in the KMeans clustering analysis in the next step, to predict the cluster associated with the certain municipality.

```
# 4. Apply k-means
kmeans = KMeans(n_clusters=3, n_init=10, random_state=0)
clusters = kmeans.fit_predict(pca_df)

# Print the cluster labels
print(clusters)
```

```
[0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 2 1 1 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0
 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]
```

Figure 3 - final labeling using k-means clustering

These labels were added to the original data frame with another label identifying if the municipality already has an IKEA. After some research into the clustering, it becomes clear what the meaning of the clustering was, the biggest municipalities are put in the 3rd cluster and medium are assigned to the 2nd one, only the smallest ones are assigned to the 1st cluster (which can also be seen in the boxplots below).

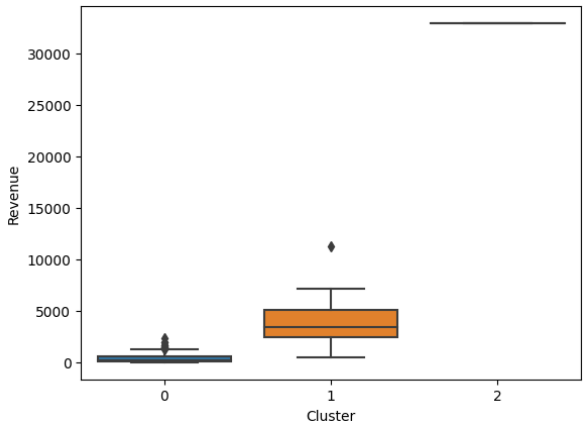

Figure 5 - boxplot of cluster label against revenue

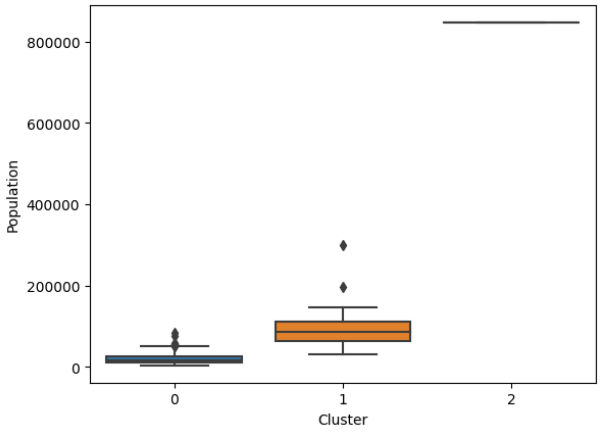

Figure 4 - boxplot of cluster label against population

Following the KMeans clustering analysis in task 1, municipalities that already had an IKEA store were excluded from the list, leaving only clusters 0 (1st cluster) and 1 (2nd cluster) for consideration.

To determine the best potential locations, the municipalities with the highest population size in each cluster were initially selected.

Cluster		Kommun_code	Year	Kommun_name	Revenue	Employee	Population	Population_Un
0	10	127	2010	Botkyrka	1469	530	82608	
	12	136	2010	Haninge	1459	1031	77054	
	133	1383	2010	Varberg	1639	1109	58084	
	90	980	2010	Gotland	1542	767	57269	
	26	188	2010	Norrtälje	1191	678	56080	
1	53	581	2010	Norrköping	4438	2630	130050	
	118	1281	2010	Lund	3438	1561	110488	
	173	1490	2010	Borås	3365	1486	103294	
	9	126	2010	Huddinge	7153	2374	97453	
	42	484	2010	Eskilstuna	2844	1485	96311	

Figure 6 - ranking of biggest kommuns per cluster

To further refine the selection process, a Score was created in task 1, which takes into account the revenue generated per person of the population. This calculation provides an indication of efficiency in potential store locations.

Using this Score, a list of the most suitable municipalities for new IKEA stores was generated.

]:

	Kommun_code	Year	Kommun_name	Revenue	Employee	Population	Population_University	Percent_Univ
98	1231	2010	Burlöv	1962	928	16701	1639	0.0
169	1486	2010	Strömstad	1281	667	11808	1071	0.0
132	1382	2010	Falkenberg	4395	2035	41008	3396	0.0
7	123	2010	Järfälla	5119	2001	66211	9009	0.1
184	1730	2010	Eda	626	271	8524	426	0.0
9	126	2010	Huddinge	7153	2374	97453	12795	0.1
115	1277	2010	Åstorp	876	326	14737	828	0.0
165	1481	2010	Mölndal	3512	1487	60973	9962	0.1
16	160	2010	Täby	3441	1546	63789	13442	0.2
103	1261	2010	Kävlinge	1455	559	29013	3911	0.1

Figure 7 - ranking of best kommun based on score

The analysis conducted in task 1 identified the top three municipalities for potential new IKEA stores as Burlöv, Strömstad, and Falkenberg. However, upon further research, it was discovered that many of these cities are located in close proximity to existing IKEA locations, rendering them unsuitable for further expansion.

After considering this additional information, the final list of top three municipalities was updated to include Strömstad, Falkenberg, and Eda. It is important to note that this research is based on several assumptions and is limited by the available data. Nonetheless, it represents a promising step towards identifying suitable locations for new IKEA stores.

Task 2: Predictive Analysis, Supervised Learning – Electricity Price Prediction

Task 2 involves predicting the daily price of electricity based on the daily consumption of heavy machinery used by businesses. To achieve this, a dataset consisting of 38,000 rows and 18 columns was loaded. The columns included information related to dates, such as week, day, month, and holidays, as well as other relevant information for research purposes.

Upon initial examination of the dataset, it was necessary to convert certain object columns to numerical columns and ignore any resulting errors. The datetime column was also converted from an object type to an actual datetime column, ensuring that all columns were in the appropriate format for correlation analysis.

Despite these initial transformations, there were still some missing values in the metric columns. To address this issue, the average of the surrounding 10 values for each missing data point was used to predict the value and provide a more accurate measurement. Any remaining rows with empty values were simply deleted from the dataset.

After this the analysis would start with looking at a correlation matrix that we could use to try and guess, which predictors could be used for the prediction of the SMPEP2 variable.

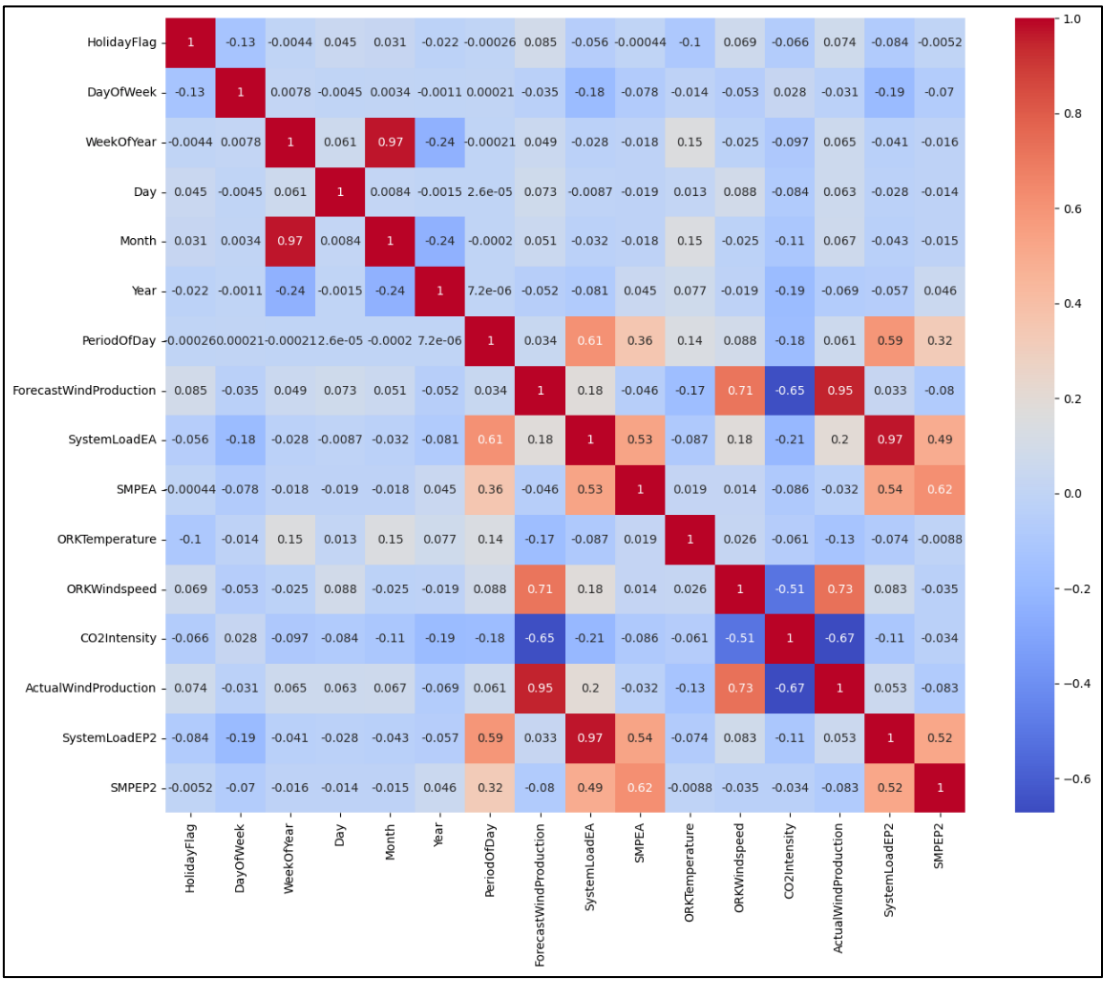

Figure 8 - correlation matrix

Based on this and the information gathered by looking at and investigating the first couple of rows of the data frame, it was quickly concluded that only 4 of the 18 columns would be close to useful and that the other ones could be thrown away for the prediction of the SMPEP2 variable. This resulted in a data frame with the following columns (in brackets behind the column is the correlation with the SMPEP2 variable):

- PeriodOfDay: half-hour period of the day (0.323050)
- SystemLoadEA: forecasted national load (0.490823)
- SystemLoadEP2: actual national system load (0.516816)
- SMPEA: forecasted national price (0.617149)
- SMPEP2: the actual price of the electricity consumed (1.000000)

Then, after this is done we started to prepare the data and split it into x and y data. So a set with the predictors and a set with only the variable that needed to be predicted (the SMPEP2 column). The we started executing different models using 10-fold cross validation:

1. Decision Tree Regression:

```
Mean MSE: 33.24729621311039
MSE standard deviation: 2.62983312533525
Mean R-squared: 0.116124172207235 74
R-squared standard deviation: 0.07536493997413814
```

2. Random Forest Regression:

```
Mean MSE: 24.240737802229233
MSE standard deviation: 2.162914257499385
Mean R-squared: 0.5311238039805433
R-squared standard deviation: 0.037913534658017486
```

3. Support Vector Regression:

```
Mean MSE: 27.070912621422515
MSE standard deviation: 2.077873122984858
Mean R-squared: 0.4147101723502987
R-squared standard deviation: 0.037905292727261945
```

These are the results of the different column numbers in combination with the different algorithms for the regression. In the end, it was chosen to not go on with the SVR algorithm, because the algorithm simply took too long to calculate and the results weren't so far not even interesting.

Then we also made some plots based on the results that were gathered, this one contains in the label table the score itself and in the first 3 tables the difference between the prediction and the actual score for all the 3 different models (based on the 3 columns):

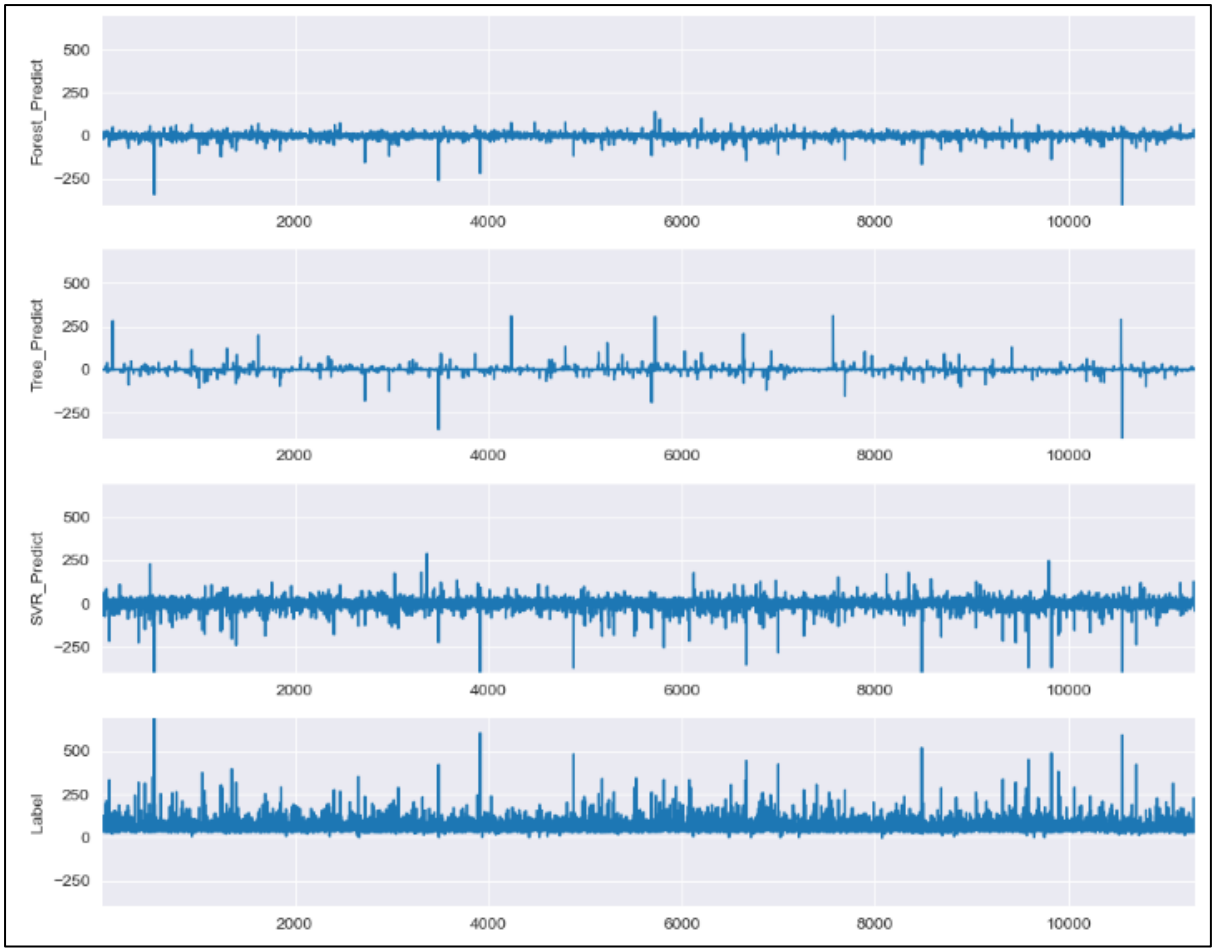

Figure 9 - all the different models and the actual label

Here you can see the different predictions separated into different plots with their difference displayed compared to the real price value. Also, there are some scatterplots that give a comparison of a perfect prediction line, where the predicted value and the real value are exactly the same and the scattered points are the actual prediction against the real prices.

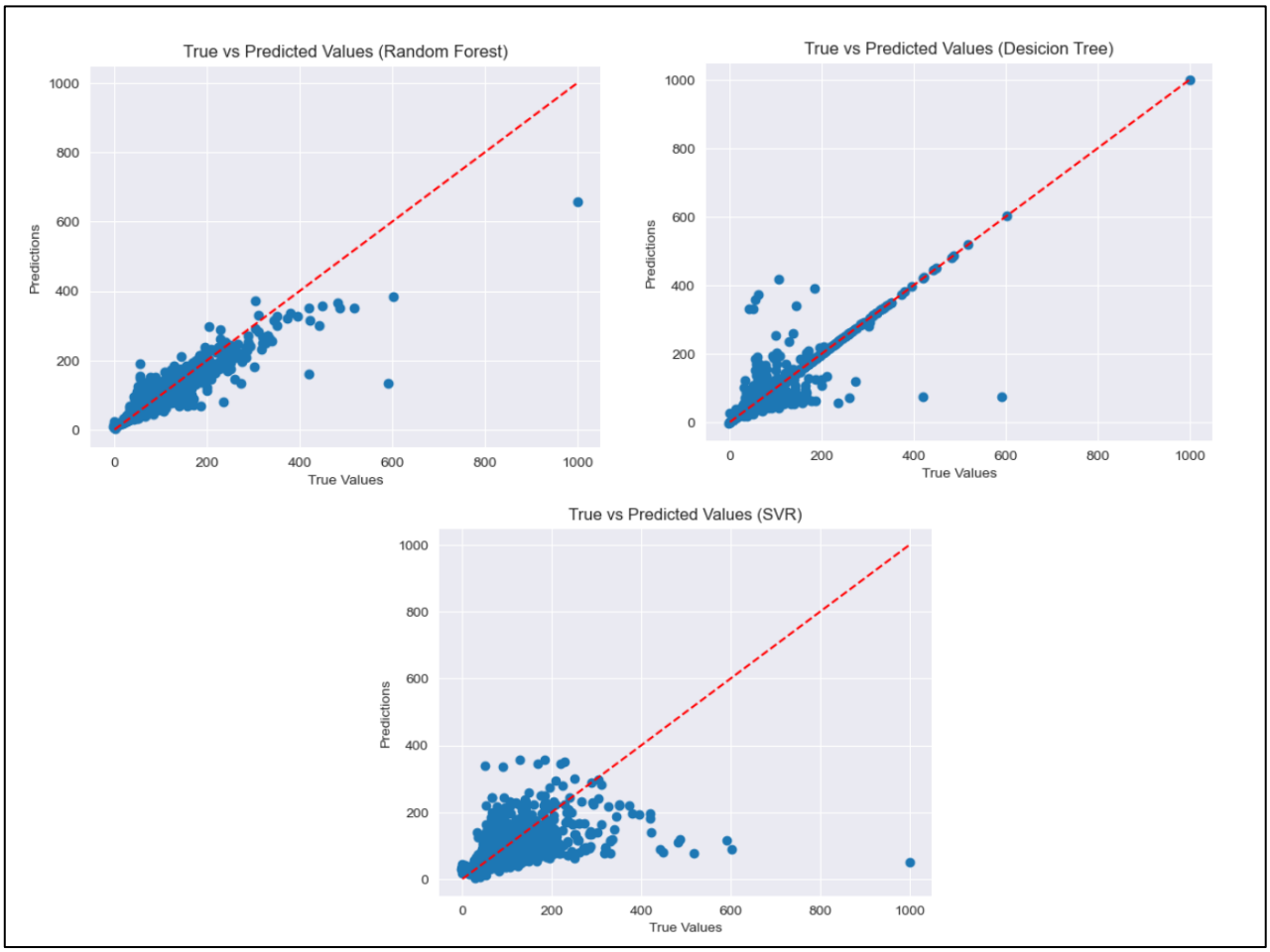

Figure 10 - models with prediction agains true values

In the end we can conclude that the SVR model was the worst model out of the 3 different regression models it had the worst MSE score and the worst R-squared rating. Then the decision tree and the random forest regressors both had quite some good predictions and I think that these two compete quite well with each other. But in the end I think the Random Forest will win, also concluding based on the MSE score and the R-squared rating. Even though the Decision Tree model was much quicker with the calculations.