DALARNA UNIVERSITY

Assignment

Home Exercise 1

Authors: Daniel Stevens (h17daste@du.se)
Omar Omran (h24omaom@du.se)
Course name: Statistical Learning
Course code: AMI22T

1. Introduction

This assignment consisted of two tasks.
Task 1 involved analysing the passing rates in grade 9 mathematics in Dalarna, both in overall trend and with regards to any correlations with other demographic and political data. In addition to the actual analysis, this task tested data cleaning: headers need to be promoted, empty cells need to be filled, columns need to be split, columns need to be transposed into rows, data types need to be changed, and so on and so forth.
Task 2 involved "assessing usability of accelerometer in detecting real-time behaviour of a certain domestic animal species", although no specific goal for the task was given. As such, we have decided to treat this as simply being a classification problem.

2. Statistical Methods

2.1. Task 1

2.1.1. Data preparation and cleaning

The data from four different sources were cleaned and merged. We transformed wide data (where each year was a column) into long data (with one row per year) to allow for easier analysis over time and missing values in the dataset were handled using forward filling for election results and population data so that we had complete data for each municipality over all years.
We removed the 2014 row because our study window is 2015–2024 and anything before that would have messed with the long trend we're tracking.

2.1.2. Exploratory Data Analysis (EDA)

We began by visualizing the correlation matrix of numeric variables to see how different variables relate to passing rates which helped us identify which variables were strongly correlated with grade 9 results and which ones were highly correlated with each other so we could avoid multicollinearity in our regression model.

2.1.3. Modeling: Time Trend and Multiple Regression

We started by fitting a simple linear regression (grade ~ year) to check for a trend in passing rates. We then extended the model by adding predictors (income, education, political shares) and used variance inflation factor (VIF) to check for multicollinearity. Variables with high VIFs (above 10) were removed from the model. The process of removing predictors with high VIFs is similar to **backward selection**, where we started with all variables and removed those that didn't add much to the model due to multicollinearity or low significance and this step helped us to improve model stability by ensuring that predictors were independent of each other leading to a more reliable model.

We used a 0.05 significance level rather than a more lenient 0.20 to limit false positives and this is the standard threshold in most regression analysis, We also performed residual diagnostics to ensure the assumptions of linear regression (linearity, normality of residuals, and constant variance) were met.

2.1.4. Model Diagnostics

We used diagnostic plots (residuals vs fitted, Q-Q plot, scale-location, leverage plots) to check the validity of our regression model. These diagnostics confirmed that the model assumptions were satisfied and that the results were reliable.

2.2. Task 2

2.2.1. Cleaning the data

The data for task 2 required less effort to clean than the datasets found in task 1. The only columns that required type changes were “Timestamp”, which we turned into a numerical Unix time value, and “Modifiers” which required conversion to factor.

We elected to remove the rows with the modifier “Missing data” as that was analogous to a “NA” row and useless to classify other data as.

The data was then normalized before being passed into the classifier.

2.2.2. Classifying the data

As the dataset consisted of 15456 observations of 53 variables with 17 classes, we decided to use a k-Nearest Neighbours algorithm to evaluate whether the dataset was sufficient to reliably determine the class of non-trained data. To split the training data, we used a standard 70:30 train-test split.

3. Results

3.1. Task 1

3.1.1. Simple Time Trend

From the simple linear regression of grade on year we found that the slope of year was negative (−0.36), meaning that the grade 9 passing rates in Dalarna decreased slightly over the last 10 years. This result is statistically significant with a p-value of 0.0228 and the confidence interval for the slope is [−0.68, −0.05] suggesting that the actual rate of decline could be anywhere between 0.05 and 0.68 % per year.

3.1.2. Multiple Regression

```
Call:
lm(formula = Grade ~ Year + Median_income + Average_income +
    Income_Inequality + Postsec_3_or_more + Mod_Party + Centre_Party +
    Liberal_Party + Christian_Democratic_Party + Green_Party +
    Social_Democratic_Party + Left_Party + Sweden_Democrats,
    data = df_merged)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8120  -2.6072   0.4668   3.0335  10.4000

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -8.845e+02  6.306e+02  -1.403  0.16330
Year                        4.441e-01  3.095e-01   1.435  0.15385
Median_income               3.074e-01  9.873e-02   3.113  0.00231 **
Average_income             -9.013e-02  5.938e-02  -1.518  0.13167
Income_Inequality           7.667e-01  6.953e-01   1.103  0.27239
Postsec_3_or_more          -1.643e-04  2.220e-04  -0.740  0.46073
Mod_Party                  -5.346e-02  1.461e-01  -0.366  0.71504
Centre_Party                1.405e-01  9.367e-02   1.500  0.13611
Liberal_Party               7.447e-01  3.114e-01   2.391  0.01832 *
Christian_Democratic_Party -2.858e-01  1.608e-01  -1.778  0.07794 .
Green_Party                 7.276e-01  2.528e-01   2.878  0.00473 **
Social_Democratic_Party    -3.551e-02  9.068e-02  -0.392  0.69605
Left_Party                 -2.823e-01  2.550e-01  -1.107  0.27038
Sweden_Democrats            1.985e-03  1.435e-01   0.014  0.98898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.857 on 121 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared:  0.311,    Adjusted R-squared:  0.2369
F-statistic: 4.201 on 13 and 121 DF,  p-value: 9.541e-06
```

Figure 1 R Outputting Regression Coefficients

After filtering the model by removing variables with high VIFs (like population and education) which lead to the final model showing that median income and green party share were the most significant predictors of grade 9 passing rates.

Median income was positively associated with better passing rates ($\beta = 0.18$, $p < 0.01$) meaning cities with higher median income be likely to have higher pass rates.

Green party support was also positively related with higher passing rates ($\beta = 0.78$, $p = 0.0023$) suggesting that cities with more green party support also had better results in mathematics. other variables like income inequality and Sweden democrats did not show significant relationships with passing rates once we oversaw for the other factors.

3.1.3. Model Diagnostics

Adjusted R^2 for the final model was 0.23, meaning that about 23% of the variation in passing rates can be explained by the model and the residual standard error was 4.88 showing the typical prediction error is about 4.88%.

Diagnostic plots confirmed that the residuals of the model met the expectations of linear regression: no non-linearity or heteroscedasticity, and no influential data points.

3.2. Task 2

Upon running the KNN classifier, we found that the elbow point for the k-value appeared to be at k=8.

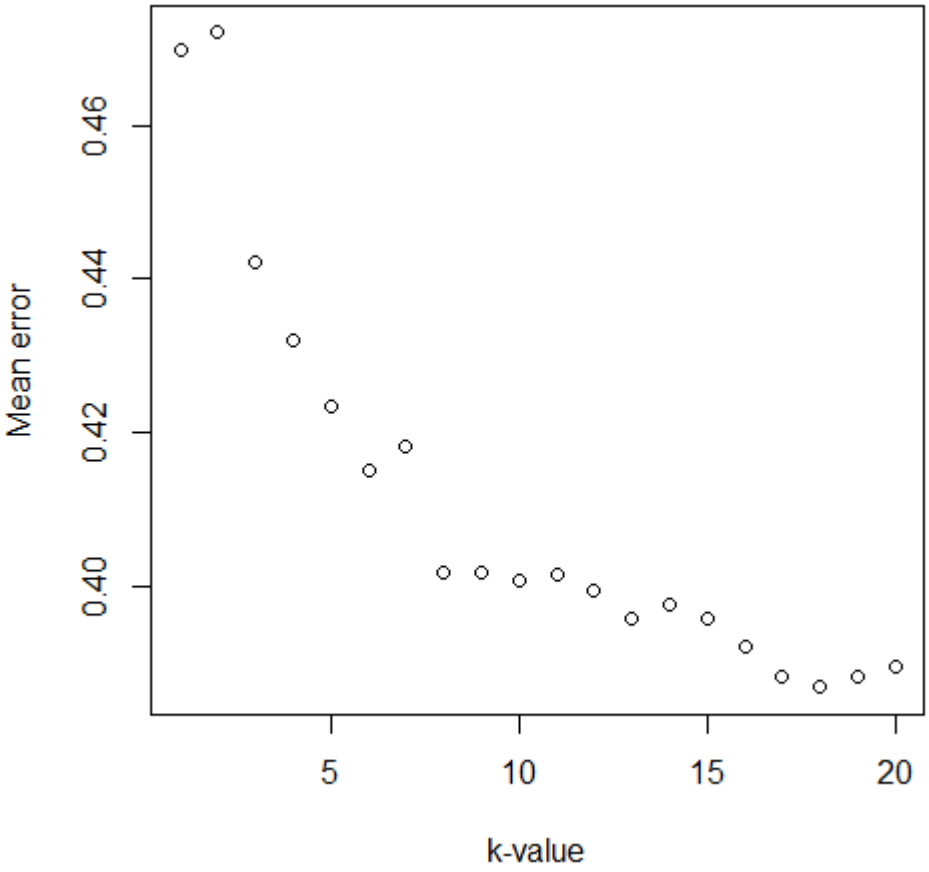

Figure 2Plot showing the mean error of increased K values

At K=8, the mean error of the model is 40.2%, or a success rate of 59.8%.

4. Discussion

4.1. Task 1

4.1.1. Overall Trend

The data shows a small but statistically significant down trend in grade 9 mathematics passing rates in Dalarna with a decline of about 0.36 percentage per year from 2015 to 2024. while this decline is relatively small it is concerning that it has carried over nearly a decade.

4.1.2. Demographics and Politics

The analysis shows that higher median income and stronger green party support are significantly connected with better passing rates and this implies that richer municipalities and those with more progressive political leanings tend to perform better on the math exam and these results show the importance of addressing inequality in education and ensuring that cities with lower income levels receive the support they need.

4.1.3. Model Strengths and Limitations

The final model explains around 23% of the variation in passing rates keep in mind while this is not a very high R^2, it is common for educational outcomes to be influenced by a wide range of complex and unmeasured factors, we took care to check multicollinearity and ensure that our model's assumptions were valid through residual diagnostics.

4.1.4. Future Research

Further research could explore non-linear effects or interactions between income, education, and political factors it would also be useful to look at whether certain educational interventions (such as targeted support in lower income areas) could help reduce the observed gap in performance between cities.

4.2. Task 2

With a 59.8% success rate, our model outperforms random guessing (5.9% success rate) by a factor of ten. However, it will still miscategorize the animal's behavior four times out of ten, which may be an unacceptable failure rate depending on the usecase.