HOME EXAMINATION

BAN404

Spring, 2023

Date: 15.05.2023

Time: 8.00-16.00

Number of hours: 8

THE HOME EXAMINATION SHOULD BE SUBMITTED IN WISEFLOW

You can find information on how to submit your paper here:
https://www.nhh.no/en/for-students/examinations/home-exams-and-assignments/

Your candidate number will be announced on StudentWeb. The candidate number should be noted on all pages (not your name or student number). In case of group examinations, the candidate numbers of all group members should be noted.

Collaboration between individuals or groups on submission preparation, as well as exchange of self-produced materials between individuals or groups is prohibited. The answer paper must consist of individual's or the group's own assessments and analysis. All communication during the home exam is considered cheating. All submitted assignments are processed in Urkund, a plagiarism control system used by NHH.

SUPPLEMENTARY REGULATIONS TO HOME EXAMINATIONS

You can find supplementary regulations under the headline "Regulations"
https://www.nhh.no/en/for-students/regulations/

Find more information under chapter 4.0 in the Supplementary provisions to the regulations for fulltime study programmes

Number of pages, including front page: 4

Number of attachments: 1, The file "Churn.csv" on Canvas.

Instructions

Your solution should be submitted as two files.

1. A pdf-file where both your R-code and the answers to the tasks are given.
2. Either a well documented R-script (.R) or an R markdown file (.Rmd) which can reproduce the results in the pdf-file.

The dataset used here has been analyzed before. Make sure to refer to other's work if you use it to solve the tasks.

A general recommendation: Be pragmatic if you run into difficulties. You need to find models that can produce predictions. Don´t let the perfect prediction stand in the way for a good one!

Each subtask is equally weighted when the exam is marked.

Throughout the exam you will work with training and test data and you will have to make appropriate choices on when to use one or the other. In cases when you think there are more than one valid choice on working with the training, test or the full dataset, explain your choice!

Tasks

In the exam you will analyze the dataset in the file `Churn.csv`, available in Canvas. This data is a modified version of the data set collected by Kunt (n.d.).

```
Churn <- read.csv("Churn.csv")
```

Customer churn is a term used for when customers exit a relationship with a company service. This data set contains information about properties of individuals who are using, or have stopped using, a service by an internet provider. In order to make decisions of potential measures to reduce churning you should predict two quantities: The size of a subscribers bill (Task 1) and whether a customer churns (Task 2).

The variables in the data set are given in the table below.

Variable	Values
`id`	Customer id
`is_tv_subscriber`	1 if Yes; 0 otherwise
`is_movie_package_subscriber`	1 if Yes; 0 otherwise
`subscription_age`	Length of subscription
`bill_avg`	Customer's bill average
`remaining_contract`	Remaining duration of contract, 0 if not applicable
`service_failure_count`	Number of experienced service failures for the customer
`download_avg`	Average downloads in GB for the customer
`upload_avg`	Average uploads in GB
`download_over_limit`	Number of times a customer exceeds the download limit
`churn`	1 if customer churn, 0 otherwise

Task 1

In this task you should predict the average bill for a subscriber.

a) If necessary, recode categorical variables as factors and remove variables that cannot be used in the analysis. Also, remove variables if you mean that they are unreasonable to include. One such reason could be that a variable is not available at the time the prediction is made but there might be other reasons too. Motivate your choices and assumptions, in words. Split the data in a 50/50 split to create a training and a test dataset. Use the seed 65923764, when you split the data, `set.seed(65923764)`.

b) Use descriptive methods to find useful predictors for `bill_avg`. Write in words which R functions you used, present the most interesting results as tables and graphs and comment on them.

c) Produce the best possible predictions of `bill_avg` using standard linear regression (OLS) and evaluate them. Motivate your choices of variables, how you evaluate the predictions and the evaluation measure used.

d) Fit a LASSO regression with all variables. Here you should use the tools you have learned to find appropriate tuning parameters. Are you standarizing the predictors or not? Motivate your choice. Compare the estimated coefficients with the estimated coefficient from an OLS regression on all predictors. (Hint: Using categorical (factor) variables in LASSO, you will have to create dummy variables of the categories. Also, note that the `glmnet`-function asks for matrices as input).

e) Compute predictions for `bill_avg` with the model fitted in 1d) and evaluate them with an appropriate measure.

f) Fit a regression tree to `bill_avg`. Explain the choices you are making. Interpret the tree.

g) Predict `bill_avg` with the regression tree in 1f) and evaluate the predictions.

h) Fit a random forest to `bill_avg`. Plot a variable importance measure for the predictors, interpret it, and, briefly, explain the measure. (If the computations are too slow, you can reduce the number of trees to grow or use a smaller training data set.)

i) Use the model in 1h) to predict `bill_avg` and evaluate the predictions.

j) f) Based on your analysis in 1a)-1i, what are the features of customers with high and, respectively, low average bills?

Task 2

In this task you should predict the churn variable, `churn`.

a) Make sure that all variables are on the right format for your analysis. Use tables and graphs and common sense to remove variables that you think will not be helpful. Motivate your choices thoroughly. Use descriptive statistics to find promising predictors for `churn`.

b) Let Y be a stochastic variable equal to 1 if a customer churn and 0 otherwise and let $p = P(Y = 1)$ be the unconditional probability to churn. The first 50 (original, not from the training or test data) observations of the variable `churn` contains observations drawn from the stochastic variable Y. Use bootstrap to compute a 95% confidence interval for p. Compare the standard approximation $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is the sample fraction of churners and n is the number of observations.

c) Fit a logistic regression with all variables to `churn`. Interpret the coefficients associated with `is_tv_subscriber` and `is_movie_package_subscriber`. For all coefficents, is the sign as you expected?

d) Use the logistic regression from 2c) to predict `churn`. Evaluate the predictions in an appropriate way. Are the predictions improved if you remove some variables from the model?

e) Use a random forest to predict `churn` and evaluate the predictions. (If the computations are too slow, you can reduce the number of trees to grow or use a smaller training data set.)

f) Based on your analysis in 2a)-2e, what are the typcial features of customers who churn?

References

Kunt, Mehmet Sabri. n.d. "ISP Data from Kaggle." https://www.kaggle.com/datasets/mehmetsabrikunt/internet-service-churn.