UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine Learning and Statistical Methods for Prediction and Classification
Day of examination:	June 9 - 2022
Examination hours:	15.00 – 19.00.

This problem set consists of 8 pages.

Appendices:	List of formulas for STK1100/STK1110 and STK2100
Permitted aids:	Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.
All subquestions are counted equally!

Problem 1

We will in this exercise look at a dataset about house prices in India[1]. After some preprocessing of the data, the following variables are available:

POST_BY Category marking who has listed the property, three categories, 'Builder', 'Dealer', 'Owner'

UND_CST Under construction or Not (0/1)

RERA Rera approved or Not (0/1)

BHK_NO Number of Rooms (1-20)

SQRFT Total area of the house in square feet

RES Category marking Resale or not (0/1)

LON Lontitude of the property

LAT Latitude of the property

Price The price on sale, the response variable

The dataset consist of 28979 properties which randomly is divided into a training set of size 14489 and a test set of size 15590.

[1] `https://www.kaggle.com/datasets/ruchi798/housing-prices-in-metropolitan-areas-of-india`

(Continued on page 2.)

(a) We start with a linear regression model using all the available explanatory variables. The table below gives the output based on fitting the training set. For categorical variables, the coefficient for the first level is put to zero.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 449.89451   27.20748   16.54  < 2e-16
POSTDealer   66.99302   10.98857    6.10  1.1e-09
POSTOwner    19.18745   11.32794    1.69  0.09032
UND_CNST1     3.71901    3.86185    0.96  0.33556
RERA1       -10.00156    3.20829   -3.12  0.00183
NoR           9.45184    2.27838    4.15  3.4e-05
SQRFT         0.11534    0.00271   42.58  < 2e-16
RESAL1      -25.30895    6.66967   -3.79  0.00015
LON          -2.83966    0.22884  -12.41  < 2e-16
LAT          -6.25635    0.33022  -18.95  < 2e-16

Residual standard error: 143 on 14479 degrees of freedom
Multiple R-squared:  0.317,     Adjusted R-squared:  0.316
F-statistic:  745 on 9 and 14479 DF,  p-value: <2e-16
```

Explain the different rows in the table for Coefficients.

In particular, discuss why there are two rows for POST_BY.

Do the output indicate that the explanatory variables have any power in explaining Price? Give arguments for your answer based on the output above.

(b) A stepwise model selection procedure based on the AIC criterion resulted in the output from the final model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 450.2186    27.2053   16.55  < 2e-16
POSTDealer   66.9530    10.9885    6.09  1.1e-09
POSTOwner    18.7581    11.3191    1.66    0.098
RERA1        -9.1399     3.0810   -2.97    0.003
NoR           9.4689     2.2783    4.16  3.3e-05
SQRFT         0.1151     0.0027   42.63  < 2e-16
RESAL1      -26.6641     6.5195   -4.09  4.3e-05
LON          -2.8469     0.2287  -12.45  < 2e-16
LAT          -6.2319     0.3292  -18.93  < 2e-16

Residual standard error: 143 on 14480 degrees of freedom
Multiple R-squared:  0.317,     Adjusted R-squared:  0.316
F-statistic:  838 on 8 and 14480 DF,  p-value: <2e-16
```

Explain why some of the estimates of the regression coefficients now have changed.

Why do you think `POSTOwner` is still included in the model even if the corresponding P-value is large?

(c) The plot below shows a fitted regression tree based on the training data.

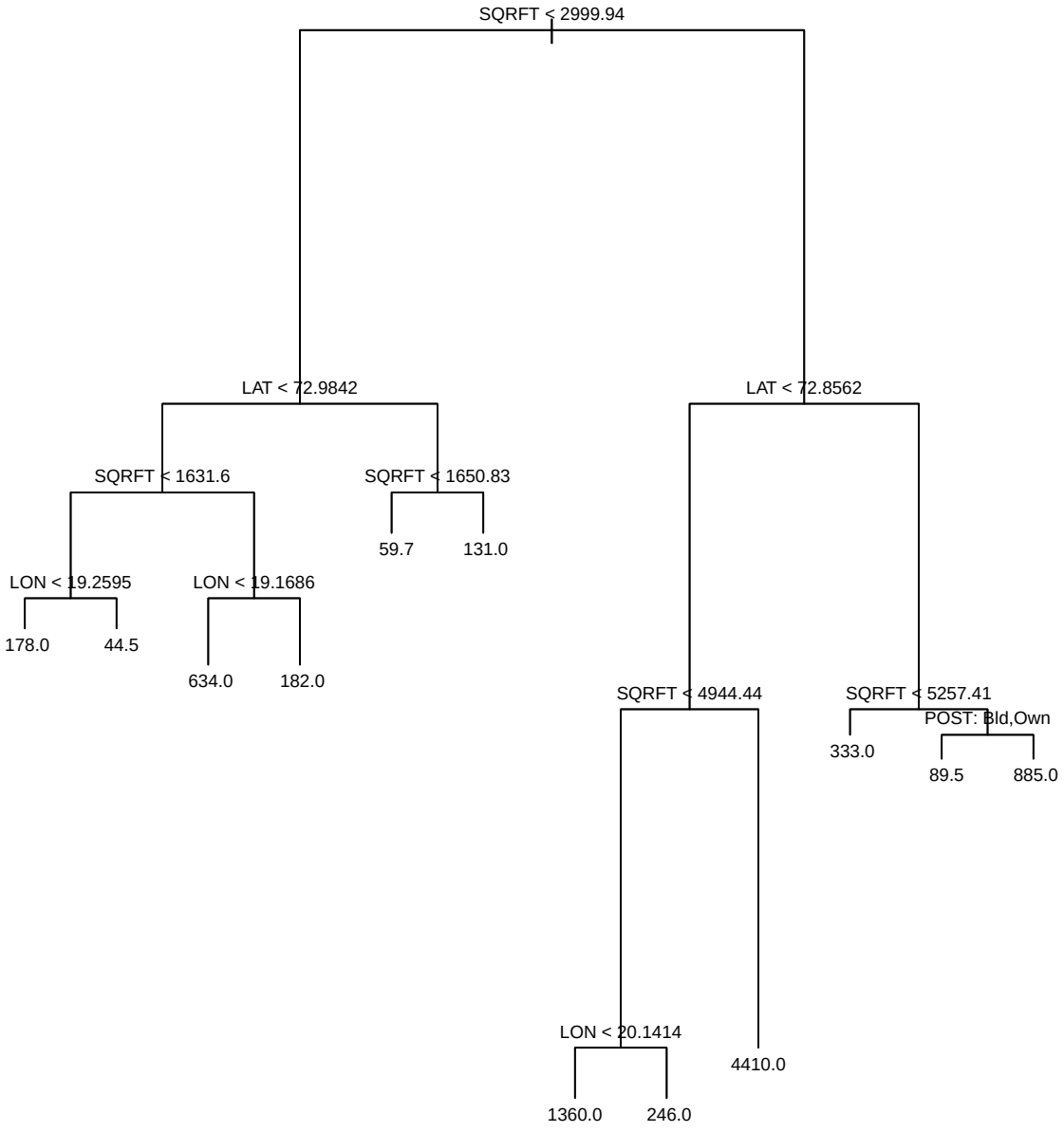
Which variables are important now?

(Continued on page 3.)

One house in the test set has the following covariates:

POST	UND_CNST	RERA	NoR	SQRFT	RESAL	LON	LAT
Dealer	0	0	3	3086.98	1	19.00	72.83

What will be the predicted value for this house?



(d) The table below give some summary results for the models/methods considered as well as some other methods. Here $RSS = \sqrt{\frac{1}{nTe}\sum_{i=1}^{nTe}(\hat{y}_i - y_i)^2}$ where the sum is taken over the *test* set.

(Continued on page 4.)

Model	Log-likelihood	RSS
Full linear model	-94013.41	158.03
Reduced linear model	-94013.88	158.03
GAM	-92617.91	148.35
TREE	-93712.78	147.20
Bagging		106.31
Random Forrest		98.66
Boosting		146.24

Based on these numbers, calculate the AIC values for the models where log-likelihood values are available. You may use here that the degrees of freedom for the GAM model is 42.4.

Both based on the AIC values and the RSS values, suggest a ranking of the different models/methods.

Problem 2

Consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$ and all noise terms assumed independent.
Based on earlier experiments one believes that β_1 is close to 1.5. Due to this, the estimation is based on the following cost function:

$$C(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda(\beta_1 - 1.5)^2.$$

(a) Define $\tilde{x}_i = x_i - \bar{x}$ and show that

$$C(\beta_0, \beta_1) = \widetilde{C}(\tilde{\beta}_0, \beta_1) = \sum_{i=1}^{N} (y_i - \tilde{\beta}_0 - \beta_1 \tilde{x}_i)^2 + \lambda(\beta_1 - 1.5))^2.$$

where now $\tilde{\beta}_0 = \beta_0 + \beta_1 \bar{x}$. Also show that $\sum_{i=1}^{N} \tilde{x}_i = 0$.

(b) Derive formulas for estimates of $\tilde{\beta}_0$ and β_1 that minimizes the cost function above. Use this to derive formulas for estimates β_0 as well.

(c) Show that for some suitable $\alpha \in [0, 1]$,

$$\hat{\beta}_1 = \alpha \hat{\beta}_1^{OLS} + (1 - \alpha) 1.5$$

where

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) y_i}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Give an interpretation of this result.

(Continued on page 5.)

Problem 3 Binary outcomes

In this exercise, we will look at the `MAGIC Gamma Telescope Data Set`[2]. The data concern registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The goal is to discriminate statistically those caused by primary gammas (signal) from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background). Thus, the response variable is whether the telescope detects one of the classes: $\{g, h\}$ standing for gamma (signal) or hadron (background). There are 10 explanatory variables, all numerical. The training data set has 9 020 observations and the test data set has 10 000 observations. The variables are as follows:

```
fLength: # major axis of ellipse [mm]
fWidth:  # minor axis of ellipse [mm]
fSize:   # 10-log of sum of content of all pixels [in #phot]
fConc:   # ratio of sum of two highest pixels over fSize [ratio]
fConc1:  # ratio of highest pixel over fSize [ratio]
fAsym:   # distance from highest pixel to center,
                        # projected on major axis [mm]
fM3Long: # 3rd root of third moment along major axis [mm]
fM3Trans:# 3rd root of third moment along minor axis [mm]
fAlpha:  # angle of major axis with vector to origin [deg]
fDist:   # distance from origin to center of ellipse [mm]
outcome: g,h # gamma (signal), hadron (background)
    n_g = 12332
    n_h = 6688
```

First, a fit to a logistic regression model gave the following result:

Coefficients (**model** 1):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.5351962	0.4529638	14.428	< 2e−16	***
fLength	−0.0300462	0.0015452	−19.445	< 2e−16	***
fWidth	−0.0027942	0.0035441	−0.788	0.4305	
fSize	−0.6422409	0.1399111	−4.590	4.43e−06	***
fConc	1.5179688	0.7595029	1.999	0.0456	*
fConc1	−7.7065391	1.1050739	−6.974	3.09e−12	***
fAsym	−0.0001170	0.0006264	−0.187	0.8519	
fM3Long	0.0075696	0.0007755	9.761	< 2e−16	***
fM3Trans	−0.0004946	0.0016491	−0.300	0.7643	
fAlpha	−0.0440680	0.0012258	−35.951	< 2e−16	***
fDist	−0.0010317	0.0004371	−2.360	0.0183	*

The log-likelihood value for this fit was **-4164.612.**

(a) Which variables increase significantly the probability of observing the signal among those analyzed in the study? How can one interpret the estimate of the regression coefficient for fLength? What about fM3Long?

Consider an alternative model with only the fLength and fAlpha covariates addressed:

[2]Dua, D. and Graff, C. (2017): UCI Machine Learning Repository`http://archive.ics.uci.edu/ml`

(Continued on page 6.)

Coefficients (**model** 2):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.3007651	0.0709471	46.52	<2e−16	***
fLength	−0.0251366	0.0008719	−28.83	<2e−16	***
fAlpha	−0.0461993	0.0011377	−40.61	<2e−16	***

The log-likelihood value for this fit was **-4337.698.**

(b) Explain why the log-likelihood value for this new model is lower than for the first model.

What are the values of AIC and BIC for the reduced and the full models? Which model would you prefer with respect to them?

Further, fitting a logistic model based on the first 2 principal components gave the following result:

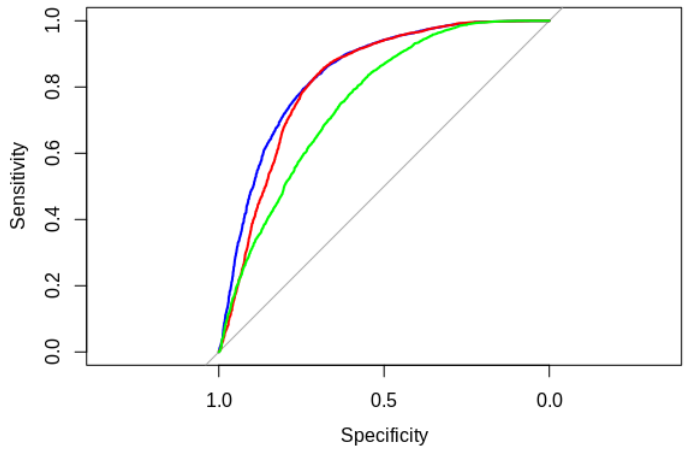
Coefficients (**model** 3):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.58565	0.02555	22.92	<2e−16	***
PC1	0.28814	0.01519	18.96	<2e−16	***
PC2	−1.11566	0.03249	−34.34	<2e−16	***

The log-likelihood value for this fit was **-4788.599.**

(c) What are the values of AIC and BIC for this model? Compare them to those previously obtained for the reduced and full model.

In the figure below, one can see the ROC curves for the three models addressed above. The three curves have areas under the curve (AUC) of **0.8392**, **0.8200**, and **0.7556** respectively.



(d) Explain what a ROC curve and an AUC are, defining accurately the concepts of sensitivity and specificity.

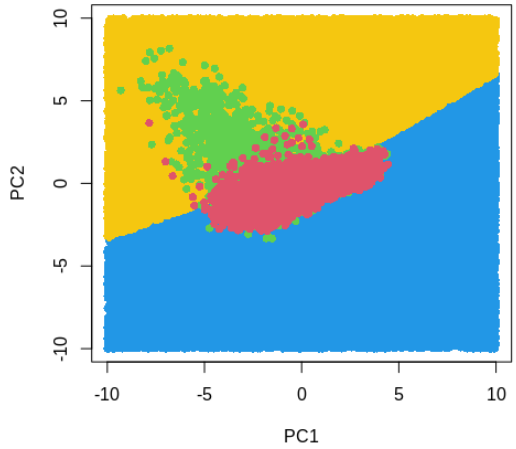
Associate the models (model 1 - model 3) with the colors of the ROC curves on a plot, explain your choice. Do the results agree with AIC and BIC?

Which model performs best for a given test set? Why?

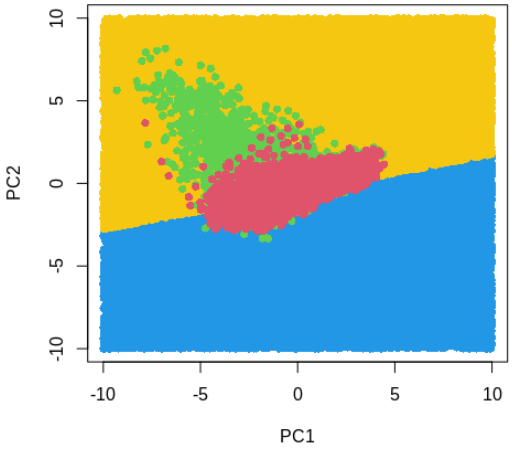
(Continued on page 7.)

Now alternative models 4-7 were fit to the data with the first two principal components used as covariates. In the figures below, the corresponding decision boundaries are shown. The respective AUC were **0.7622**, **0.7577**, **0.7502**, **0.7386** for models 4, 5, 6 and 7. Models 4-7 include (in some order) a classification tree (CT), K nearest neighbors approach (KNN), linear discriminant analysis approach (LDA) and an artificial neural network (ANN).

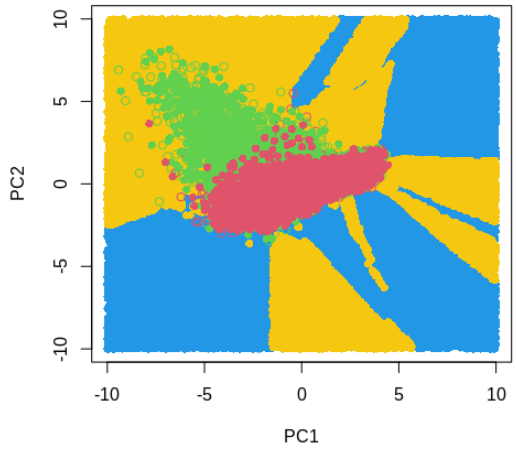
Model 4



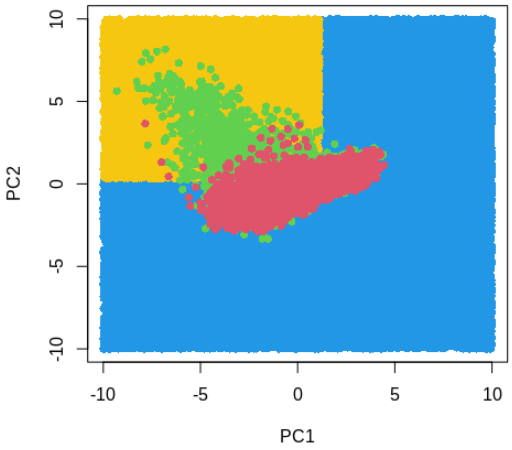
Model 5



Model 6



Model 7



(e) Associate the decision boundaries and the corresponding AUCs for models 4-7 with the models in set ANN, KNN (with 5 nearest neighbors), LDA, CT. Note that ANN is fit with strong regularization and KNN is based on 5 nearest neighbors. Explain your choice.

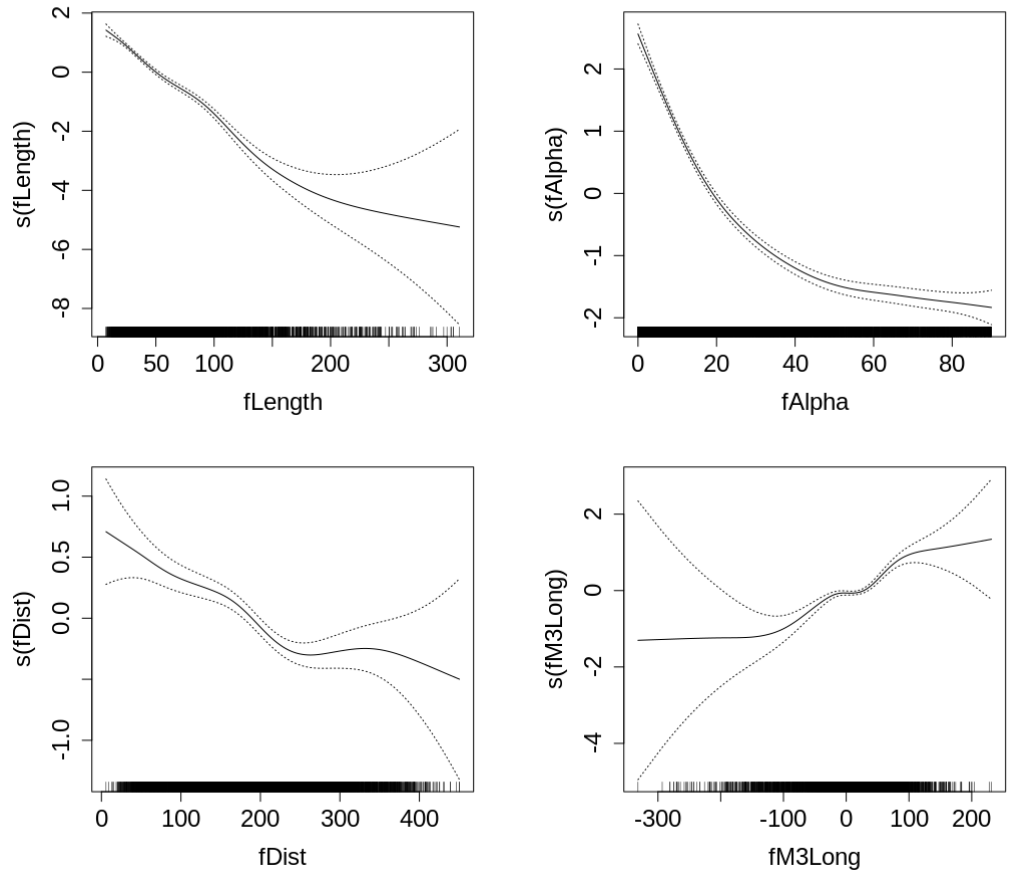
Which model from models 1-7 performs best at prediction with respect to AUC on a given test set?

The following generalised additive model (GAM) has been fitted to the data

$$p(y) = \text{logit}^{-1}(s(fLength) + s(fAlpha) + s(fDist) + s(fM3Long)).$$

The model gave an AUC of 0.8588 for the addressed test set. The figure below shows the GAM plot for the addressed covariates.

(Continued on page 8.)



Below is the summary of the GAM model

```
Anova for Parametric Effects
             Df  Sum Sq Mean Sq  F value    Pr(>F)
s(fLength)    1   249.4  249.36  179.911 < 2.2e-16 ***
s(fAlpha)     1  1479.1 1479.10 1067.169 < 2.2e-16 ***
s(fDist)      1    62.4   62.39   45.014 2.074e-11 ***
s(fM3Long)    1    95.6   95.56   68.944 < 2.2e-16 ***

Anova for Nonparametric Effects
            Npar Df Npar Chisq    P(Chi)
(Intercept)
s(fLength)        3      23.32 3.469e-05 ***
s(fAlpha)         3     452.51 < 2.2e-16 ***
s(fDist)          3      18.62 0.0003275 ***
s(fM3Long)        3      53.75 1.270e-11 ***
```

(f) Give a short interpretation of the estimated effect of each of the four explanatory variables, i.e. describe how they may affect the probability of observing the signal.

Discuss if the non-linearities are important here.

END