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 5 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

(a) Each row gives estimates for the regression coefficients involved in the model. For numerical variables we just get one corresponding regression coeficient. For categorical variables we get one regression coefficient for each level, except for the first level. This is why we get two rows for POST.

Due to that the F-statistic is very high with a corresponding very low P-value, the null hypthesis on that all the regression coefficients are zero is clearly rejected, indicating that the explanatory variables indeed have explanatory power.

(b) The regression coefficients are based on the contribution conditional on the other covariates involved. If there are correlations between the covariates, removing one variable can effect the contribution from other variables.

`POSTOwner` is still included because the selection procedure look at the whole covariate POST (including all levels) when considering this variable and since Dealer is significantly different from Builder, the covariate is kept.

(c) Now SQRFT, LAT, LON and POST are the only variables that seem important.

For this specific choice, the predicted value will be 1360.

(Continued on page 2.)

(d) Full linear model

$$AIC = -2 * (-94013.41) + 2 * 11 = 188048.82$$

For the reduced linear model

$$AIC = -2 * (-94013.88) + 2 * 10 = 188047.76$$

For GAM, we have

$$AIC = -2 * (-92617.91) + 2 * 42.4 = 185320.62$$

For TREE, we have 12 terminal nodes which correspond to the degrees of freedom, giving

$$AIC = -2 * (-93712.78) + 2 * 12 = 187449.56$$

Among these four methods, GAM has the lowest AIC value. Note however that TREE has a better prediction accuracy on the test set. Based on that combined with the easier interpretation of TREE, a ranking btween these would be TREE, GAM, Reduced LM, full LM.

If prediction is the main issue, all the other four methods (except boosting) do much better. Given the more difficult interpretation of Boosting, I would rank these Random Forrest, Bagging, TREE, GAM, Boosting, Reduced LM, full LM.

Problem 2

(a) We have

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

That $\sum_{i=1}^n \tilde{x}_i = 0$ follows from the definition.

(b) We have

$$\begin{aligned}
\frac{\partial}{\partial\tilde{\beta}_0}\widetilde{C}(\tilde{\beta}_0,\beta_1) =& -2\sum_{i=1}^N (y_i - \tilde{\beta}_0 - \beta_1\tilde{x}_i)\\
=& -2\sum_{i=1}^N (y_i - \tilde{\beta}_0)
\end{aligned}$$

which, put to zero gives $\hat{\hat{\beta}}_0 = \bar{y}$. Further,

$$\begin{aligned}\frac{\partial}{\partial\beta_1}\widetilde{C}(\bar{y},\beta_1) =& -2\sum_{i=1}^N (y_i - \bar{y} - \beta_1\tilde{x}_i)\tilde{x}_i + 2\lambda(\beta_1 - 1.5)\\ =& -2\sum_{i=1}^N (y_i - \beta_1\tilde{x}_i)\tilde{x}_i + 2\lambda(\beta_1 - 1.5)\end{aligned}$$

which, put to zero, gives

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^N \tilde{x}_i y_i + 1.5\lambda}{\sum_{i=1}^N \tilde{x}_i^2 + \lambda}\\ =& \frac{\sum_{i=1}^N (x_i - \bar{x})y_i + 1.5\lambda}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}\end{aligned}$$

We then also get

$$\hat{\beta}_0 = \bar{y} + \bar{x}\frac{\sum_{i=1}^N (x_i - \bar{x})y_i + 1.5\lambda}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}$$

(c) We have

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^N \tilde{x}_i y_i + 1.5\lambda}{\sum_{i=1}^N \tilde{x}_i^2 + \lambda}\\ =& \frac{\sum_{i=1}^N (x_i - \bar{x})y_i + 1.5\lambda}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}\\ =& \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}\cdot\frac{\sum_{i=1}^N (x_i - \bar{x})y_i}{\sum_{i=1}^N (x_i - \bar{x})^2} + \frac{\lambda}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}1.5\\ =& \alpha\hat{\beta}_1^{OLS} + (1-\alpha)1.5\end{aligned}$$

with

$$\alpha = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2 + \lambda}.$$

This shows that the estimate of $\hat{\beta}_1$ is a weighted sum of $\hat{\beta}_1^{OLS}$ and 1.5, which is reasonable given the two sources of information that we have here. Note also that α will tend to 1 as N increases.

Problem 3

(a) All covariates but fWidth, fAsym, and fM3Trans are statistifically significant on a 0.05 level of significance. Of them, fConc and fM3Long increase the probability of observing signal due to positive effects.

(Continued on page 4.)

Regarding the second part of the question, when fLength increases by one unit and other covariates are fixed the odds ratio is 0.9704 as compared to the baseline value, which reduces the probability of observing signal. When fM3Long increases by one unit and other covariates are fixed the odds ratio is 1.0076 as compared to the baseline value, which slightly increases the probability of observing signal.

(b) For the full model we have 10 covariates leading to AIC of $-2 * (-4164.612) + 11 * 2 = 8351.224$ and given the size of the training data of 9020, the BIC becomes $-2*(-4164.612)+11*\log(9020) = 8429.403$. The likelihood of the reduced model will always be smaller or equal that the one of the full model due to reduced number of degrees of freedom and hence reduced flexibility. For the reduced model we have only 2 covariates leading to AIC of $-2 * (-4337.698) + 3 * 2 = 8681.396$ and BIC of $-2 * (-4337.698) + 3 * \log(9020) = 8702.718$. For both of the criteria, the full model is still preferred when corrected for the number of covariates.

(c) For the model with 2 covariates corresponding to the first 2 principal components we have an AIC of $-2 * (-4788.599) + 3 * 2 = 9583.199$ and BIC of $-2 * (-4788.599) + 3 * \log(9020) = 9604.52$. For both of the criteria, the full model is still the best one.

(d) The ROC curve is a graphical tool to visualize the performance of a classification model, and it displays the sensitivity and (1 minus) specificity when moving the threshold used to discriminate between the two response classes.

Sensitivity = true positive / (true positive + false negative), where true positive are the observations correctly identified as positive by the model, and false negative the observations incorrectly classified as negative by the model.

Specificity = true negative / (false positive + true negative), where true negative are the observations correctly identified as negative by the model, and false positive the observations incorrectly classified as positive by the model.

The Area Under the Curve (AUC) is the area under the ROC curve measure of the ability of a classifier to distinguish between classes acrooss all tresholds and is used as a summary of the ROC curve. The closer ROC curve to the left top corner the better and correspondingly the AUC becomes closer to 1. Thus the higher AUC - the better.

In our case, the highest AUC corresponds to model 1, the second highest - to model 2 and the lowest - to model 3. Hence, model 1 corresponds to the blue ROC curve, model 2 - to the red ROC curve, and model 3 - to the green ROC curve.

(Continued on page 5.)

All three criteria (AUC, AIC, BIC) agree and range model 1 to be the best, model 2 - second best and model 3 - the worst. Model one should be thus preferred.

(e) Model 6 has a very flexible decision boundary with islands which corresponds to a KNN approach with few nearest neighbours used. Model 5 is has the only linear decision boundary, which corresponds to a LDA approach, which can only have a linear decision boundary. Model 7 has a rectamngular decision boundary, which is induced by a CT due to it making leaves based on splits of the covariates. Model 4 has a slightly non-linear decision boundary, which may only correspond to ANN with enough regularisation.

According to AUC, the full logistic regression (model 1) with the AUC of 0.8392 is the best performing model for a given test set.

(f) From both the figures and the ANOVA test, we see a strong indication of significant non-linear effects for all 4 of the addressed covariates. The same is indirectly shown by AUC, which is better than the one of the full logistic regression. As for the individual effects, we see that generally higher values of fLength, fAlpha, and fDist correspond to lower probabilities of observing the signal, whilst higher values of fM3Long - to higher probabilities the relations are clearly non-linear. However the trends of that fLength, fAlpha, and fDist correspond to lower probabilities of observing the signal, whilst higher values of fM3Long - to higher probabilities agrees to the one from the full logistic regression.