INDR 450 Homework 3 Report

Umur Berkay Karakaş
69075

Question 1

a) The results of the full regression model can be seen in Figure 2.

b) R^2 is 0.383 and adjusted R^2 is 0.283, both of which are too low. So, our model and our regressors are not good enough to explain the data. Out of all regressors, only December dummy is significant with p-value of 0.00 and January dummy is almost significant with a p-value of 0.052.

```
MAPE of the full regression model on training set: 5.75141871186084
RMSE of the full regression model on training set: 717.2574837026933
MSE of the full regression model on training set: 514458.29792751936
```

c)
```
MAPE of the full regression model on test set: 23.395918099579795
RMSE of the full regression model on test set: 5005.5396738917
MSE of the full regression model on test set: 25055427.426903825
```

The error differences between training set and test set are huge. MAPE of test set is over 4 times of the MAPE for training set, RMSE of test set is almost 7 times of the RMSE of training set. This is caused by overfitting because we used too many predictors in our regression model, therefore data overfitted to training set excessively and failed to predict the test set as well enough it predicted the training set.

d) The results of the reduced regression model can be seen in Figure 4.

e) R^2 is 0.307 and adjusted R^2 is 0.284, both of which are too low. So, our model and our regressors are not good enough to explain the data. But adjusted R^2 for reduced model is higher than the full model, which means the variable except trend, January dummy and December dummy are redundant. In the reduced model, January and December dummies are significant but trend variable is insignificant. Trend variable has been insignificant in both models, and it is reasonable because the data has no trend.

```
MAPE of the reduced regression model on training set: 6.358164058264865
RMSE of the reduced regression model on training set: 763.4325181478738
MSE of the reduced regression model on training set: 582829.2097656036

MAPE of the reduced regression model on test set: 5.996805470158502
RMSE of the reduced regression model on test set: 1426.942120277627
MSE of the reduced regression model on test set: 2036163.8146224099
```

f) Considering the adjusted R^2 values, both models are equally good to explain the variability of the data. In terms of training set errors, full model has lower error values, and it is what it should be because there was an overfitting in the first model. In terms of test set errors, reduced model has significantly less error values because we got rid of the overfitting by removing redundant variables. Considering everything we got, reduced model would be more preferable.

Question 2

a) The results of the full logistic regression model can be seen in Figure 6.

b) Lag1 and lag2 are significant with p-values less than 0.05. Lag3 has a p-value of 0.07, its significance depends on significance level (significant if significance level is 0.1, insignificant if significance level 0.01 or 0.05). Jan and Dec dummies are insignificant with p-values of 0.999. They probably have marginal effect on the predictions because of having lag1, lag2 and lag3 as predictors. Also, they only have meanings for January and December months but there are 12 months in a year, so it only helps the model to predict January and December months in a sense.

c) I have done it in the code.

d) Confusion matrix for training set can be seen in Figure 8. Recall is 0.81, precision is 0.77 and F1 score is 0.77. It could be considered that the error is low.

e) Confusion matrix for test set can be seen in Figure 10. Recall is 0.92 where 35 out of 38 positive datapoints are predicted correctly. Precision is 0.66 and F1 score is 0.77. In terms of positive labeled data points, our model looks great, and it can be considered good in general.

ROC curve for test set can be seen in Figure 9. It has an AUC score of 0.74, which can be considered good as well.

f) The results of the full logistic regression model can be seen in Figure 11.

g) AUC score for the full model on the test set is 0.74, and it is 0.69 for the reduced model. Both looks fine and explaining the data to some extent since they are higher than 0.5.

Question 3

a) ROC curve and classification matrix for k = 5 can be seen in Figure 16 and 17. AUC score is 0.54 and by looking at the classification matrix, KNN classifier with k = 5 doesn't look promising since it has predicted nearly half of each positive and negative data points correctly. AUC score being 0.54 tells us the same story.

b) All relevant figures can be seen in from Figure 18 to 23. It is clear that KNN classifier is not appropriate for our dataset since the maximum AUC score we get from different k values is 0.57 and it's not good enough.

Appendix

Figure 1: Time series plot of Toyota sales

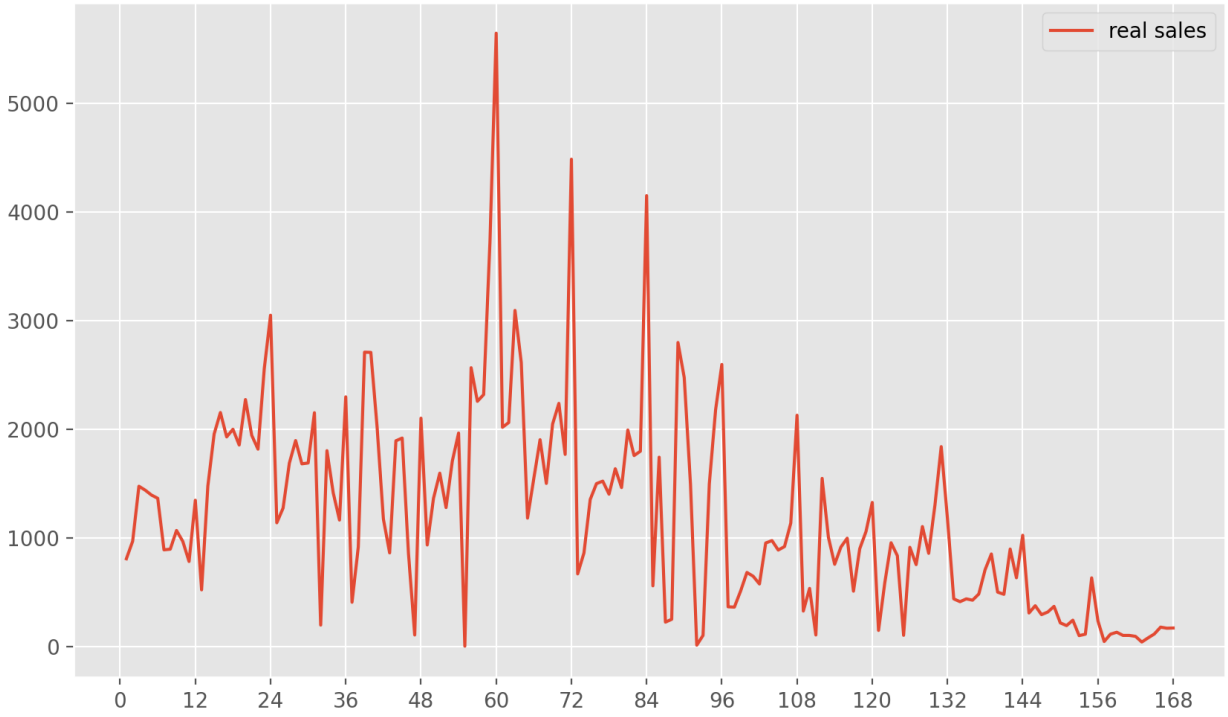

Figure 2: OLS Regression results on Toyota sales data using training set

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.388
Model:	OLS	**Adj. R-squared:**	0.283
Method:	Least Squares	**F-statistic:**	3.672
Date:	Wed, 20 Apr 2022	**Prob (F-statistic):**	9.95e-05
Time:	13:15:38	**Log-Likelihood:**	-767.46
No. Observations:	96	**AIC:**	1565.
Df Residuals:	81	**BIC:**	1603.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1369.6970	437.204	3.133	0.002	499.798	2239.596
t	-4.3960	29.980	-0.147	0.884	-64.048	55.256
tsq	0.5768	0.716	0.805	0.423	-0.848	2.002
tcube	-0.0058	0.005	-1.194	0.236	-0.015	0.004
Jan	-774.0124	392.765	-1.971	0.052	-1555.491	7.466
Feb	-322.7028	392.193	-0.823	0.413	-1103.044	457.639
Mar	106.1099	391.712	0.271	0.787	-673.275	885.494
Apr	75.3355	391.314	0.193	0.848	-703.257	853.928
May	126.2587	390.994	0.323	0.748	-651.696	904.213
Jun	51.2895	390.746	0.131	0.896	-726.172	828.751
Jul	-298.7874	390.569	-0.765	0.446	-1075.897	478.322
Aug	-294.6872	390.462	-0.755	0.453	-1071.584	482.209
Oct	-28.5661	390.463	-0.073	0.942	-805.465	748.333
Nov	126.6494	390.578	0.324	0.747	-650.479	903.778
Dec	1581.3062	390.778	4.047	0.000	803.781	2358.832

Omnibus:	1.205	**Durbin-Watson:**	1.243
Prob(Omnibus):	0.547	**Jarque-Bera (JB):**	0.684
Skew:	-0.094	**Prob(JB):**	0.710
Kurtosis:	3.368	**Cond. No.**	4.28e+06

Figure 3: Time series plot of real sales vs prediction of full regression model on the entire data

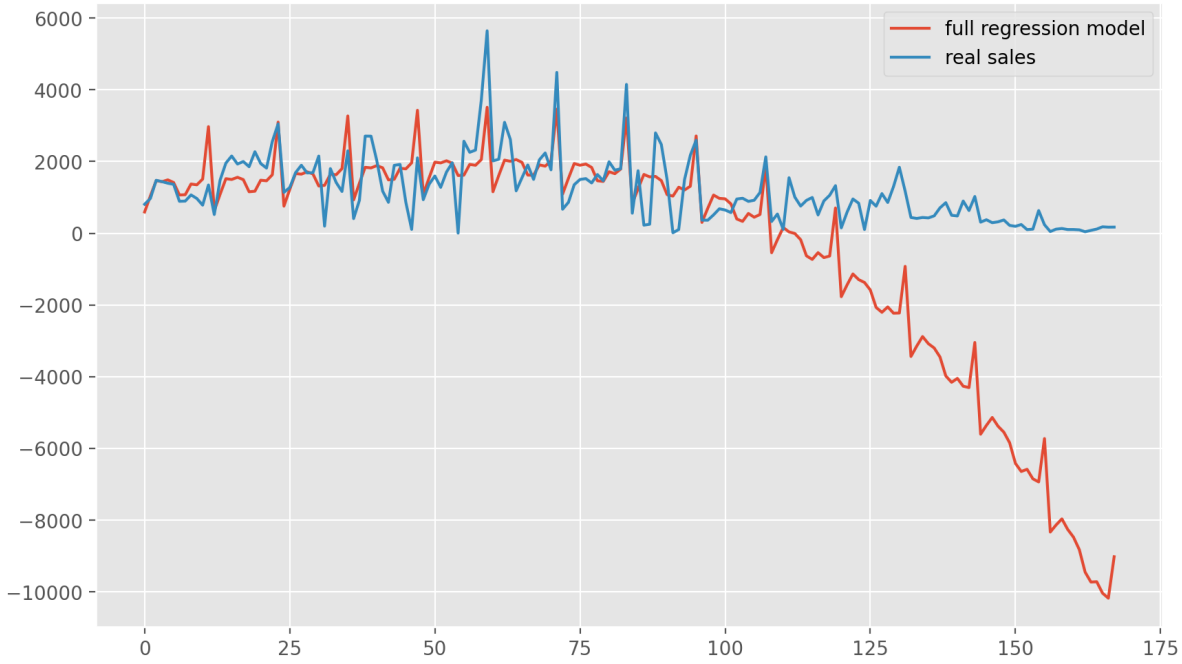

Figure 4: OLS Regression results on Toyota sales data using training set with reduced model

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.307
Model:	OLS	**Adj. R-squared:**	0.284
Method:	Least Squares	**F-statistic:**	13.58
Date:	Wed, 20 Apr 2022	**Prob (F-statistic):**	2.07e-07
Time:	13:15:38	**Log-Likelihood:**	-773.45
No. Observations:	96	**AIC:**	1555.
Df Residuals:	92	**BIC:**	1565.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1470.7058	164.729	8.928	0.000	1143.539	1797.873
t	2.6893	2.882	0.933	0.353	-3.034	8.413
Jan	-704.2213	289.611	-2.432	0.017	-1279.414	-129.029
Dec	1591.5713	289.611	5.496	0.000	1016.379	2166.764

Omnibus:	2.919	**Durbin-Watson:**	1.172
Prob(Omnibus):	0.232	**Jarque-Bera (JB):**	2.795
Skew:	0.083	**Prob(JB):**	0.247
Kurtosis:	3.819	**Cond. No.**	214.

Figure 5: Time series plot of real sales vs prediction of reduced regression model on the entire data

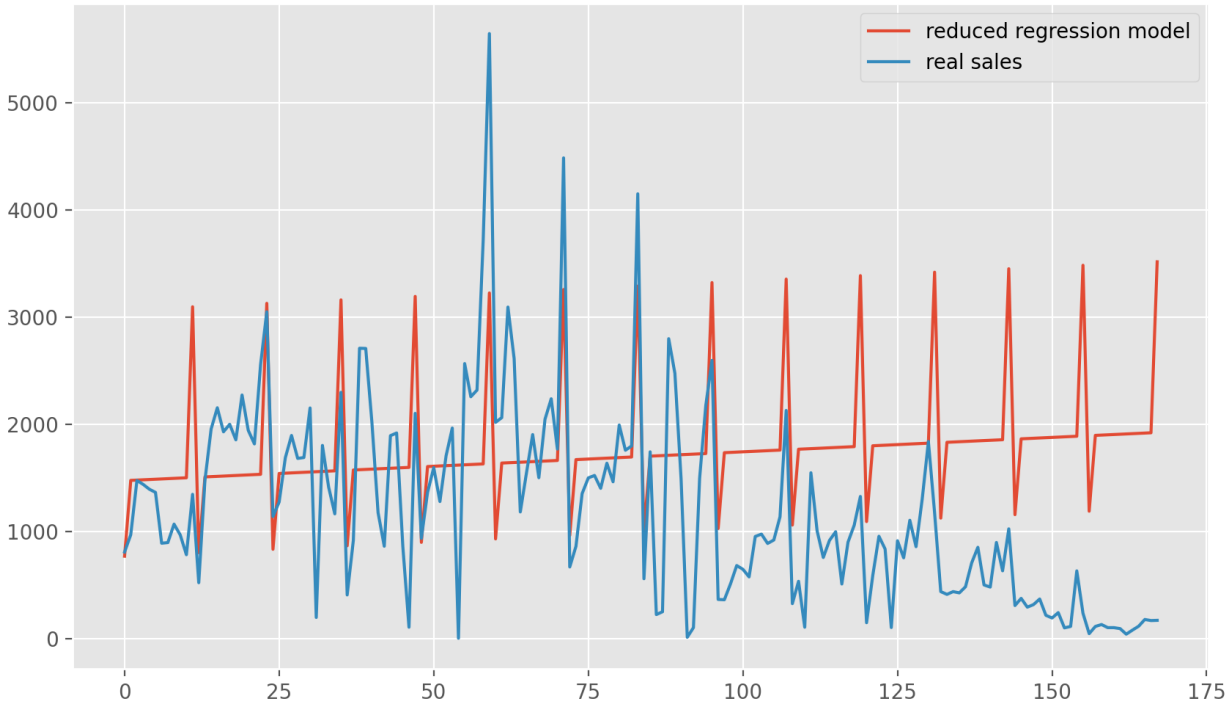

Figure 6: Full logistic regression model results on Toyota sales up training set

Generalized Linear Model Regression Results

Dep. Variable:	Up	**No. Observations:**	92
Model:	GLM	**Df Residuals:**	86
Model Family:	Binomial	**Df Model:**	5
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-43.315
Date:	Tue, 26 Apr 2022	**Deviance:**	86.630
Time:	13:09:55	**Pearson chi2:**	71.2
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.3479
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.2034	0.260	0.783	0.434	-0.306	0.713
Lag1	-0.0014	0.000	-3.347	0.001	-0.002	-0.001
Lag2	-0.0009	0.000	-2.653	0.008	-0.002	-0.000
Lag3	-0.0005	0.000	-1.813	0.070	-0.001	4.2e-05
Jan	-21.9276	2.66e+04	-0.001	0.999	-5.21e+04	5.2e+04
Dec	24.1444	2.47e+04	0.001	0.999	-4.84e+04	4.85e+04

Figure 7: ROC curve and AUC score of full logistic regression model for training set

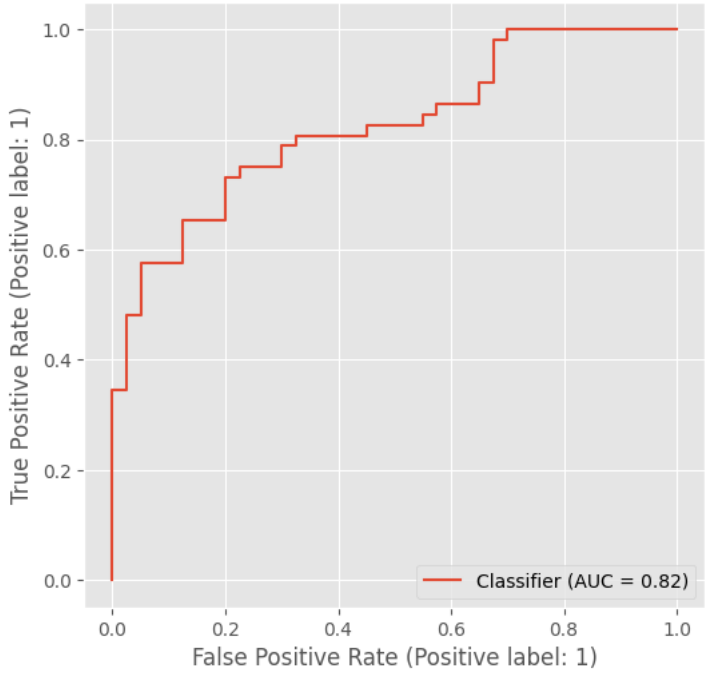

Figure 8: Confusion matrix for the training set of full logistic regression model

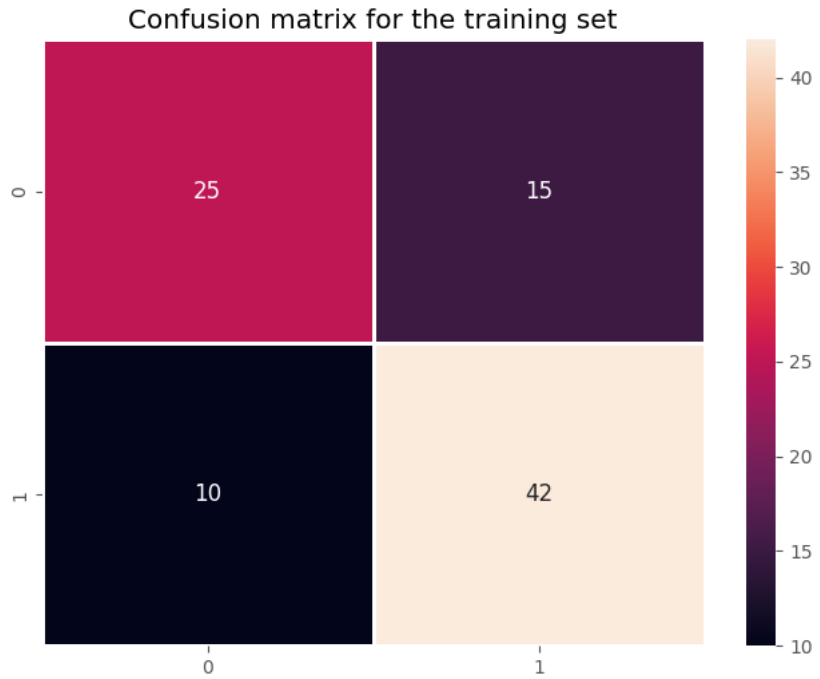

Figure 9: ROC curve and AUC score of full logistic regression model for test set

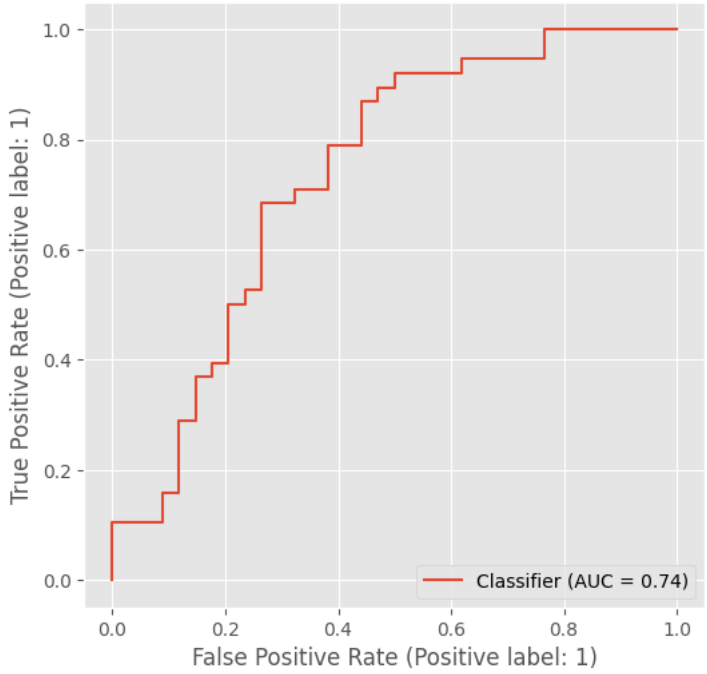

Figure 10: Confusion matrix for the test set of full logistic regression model

Figure 11: Reduced logistic regression model results on Toyota sales up training set

Generalized Linear Model Regression Results

Dep. Variable:	Up	**No. Observations:**	93
Model:	GLM	**Df Residuals:**	89
Model Family:	Binomial	**Df Model:**	3
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-47.130
Date:	Tue, 26 Apr 2022	**Deviance:**	94.260
Time:	13:21:06	**Pearson chi2:**	80.1
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.3014
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.1175	0.250	0.469	0.639	-0.373	0.608
Lag1	-0.0013	0.000	-3.628	0.000	-0.002	-0.001
Lag2	-0.0006	0.000	-2.358	0.018	-0.001	-0.000
Dec	22.9878	1.51e+04	0.002	0.999	-2.95e+04	2.95e+04

Figure 12: ROC curve and AUC score of reduced logistic regression model for training set

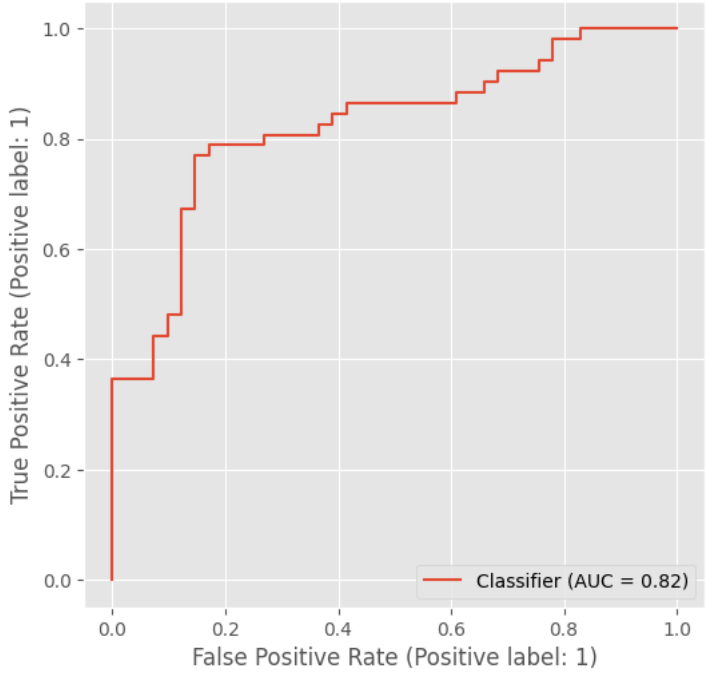

Figure 13: Confusion matrix for the training set of reduced logistic regression model

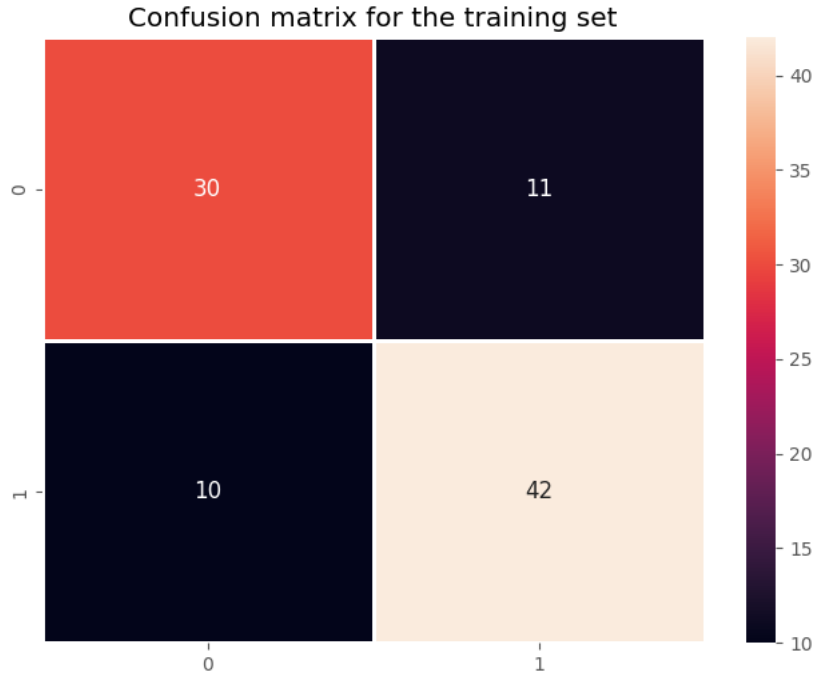

Figure 14: ROC curve and AUC score of reduced logistic regression model for test set

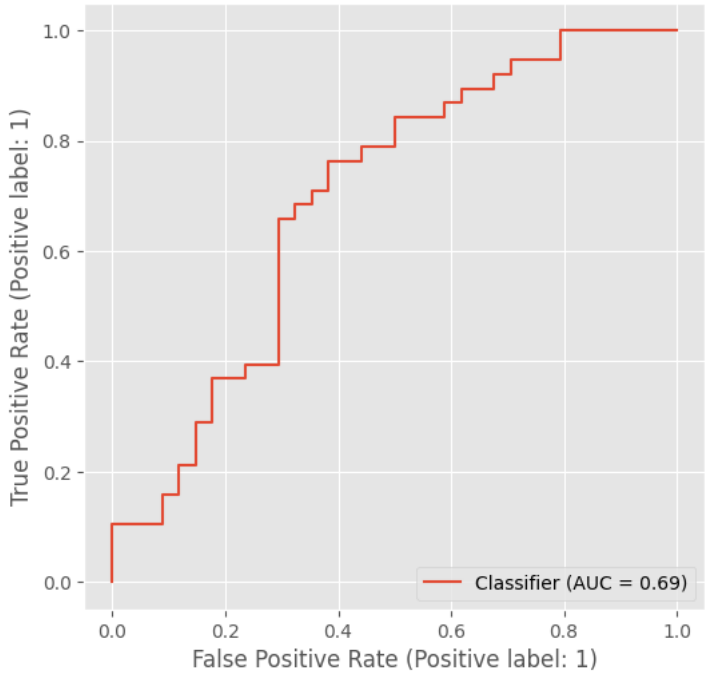

Figure 15: Confusion matrix for the test set of reduced logistic regression model

Figure 16: ROC curve and AUC score of KNN classifier for k=5

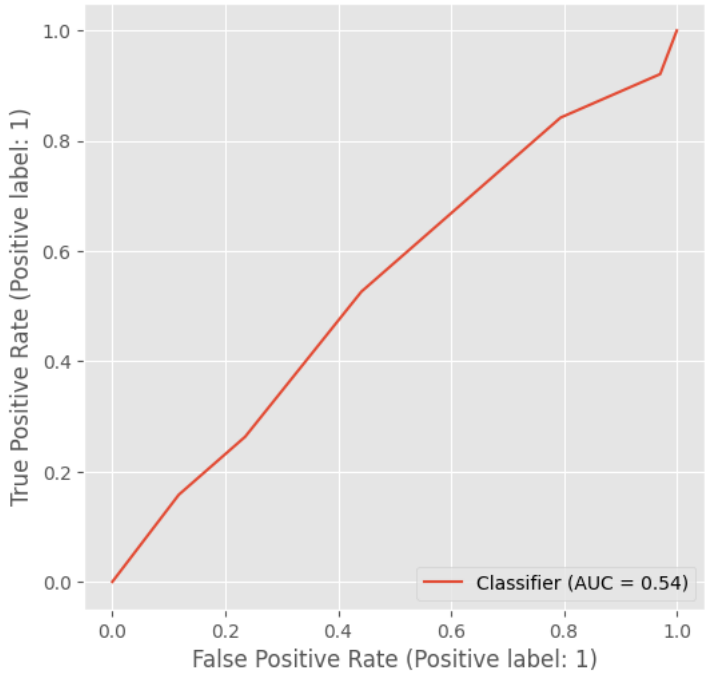

Figure 17: Confusion matrix for KNN classifier for k=5 on the test set

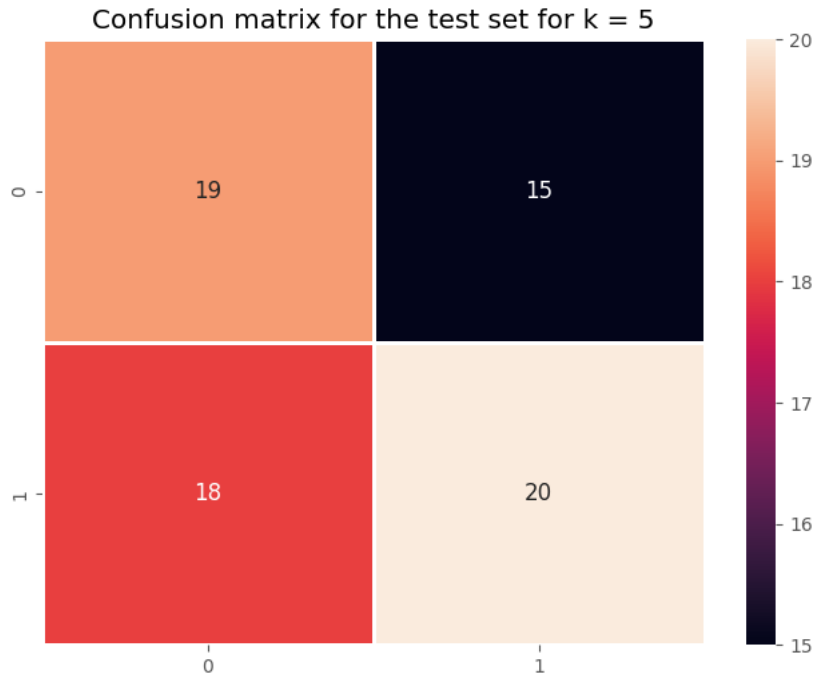

Figure 18: ROC curve and AUC score of KNN classifier for k=1

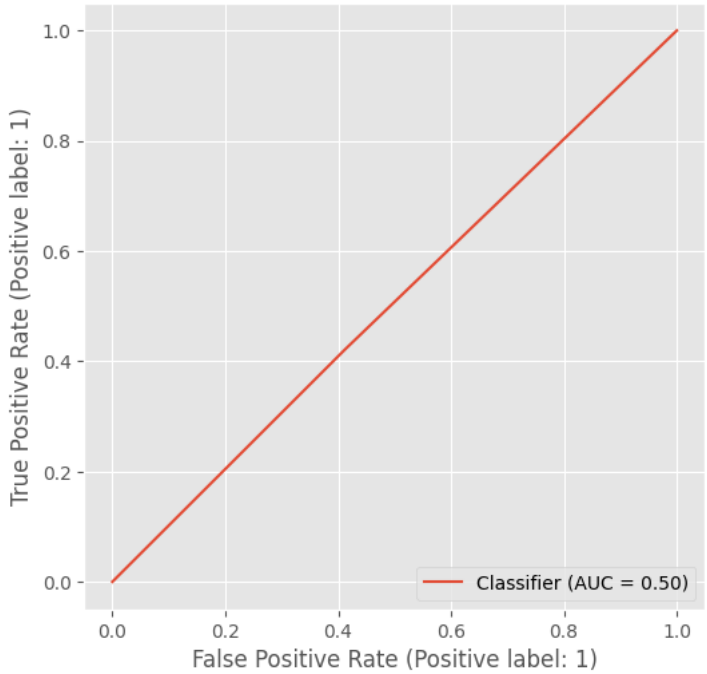

Figure 19: Confusion matrix for KNN classifier for k=11 on the test set

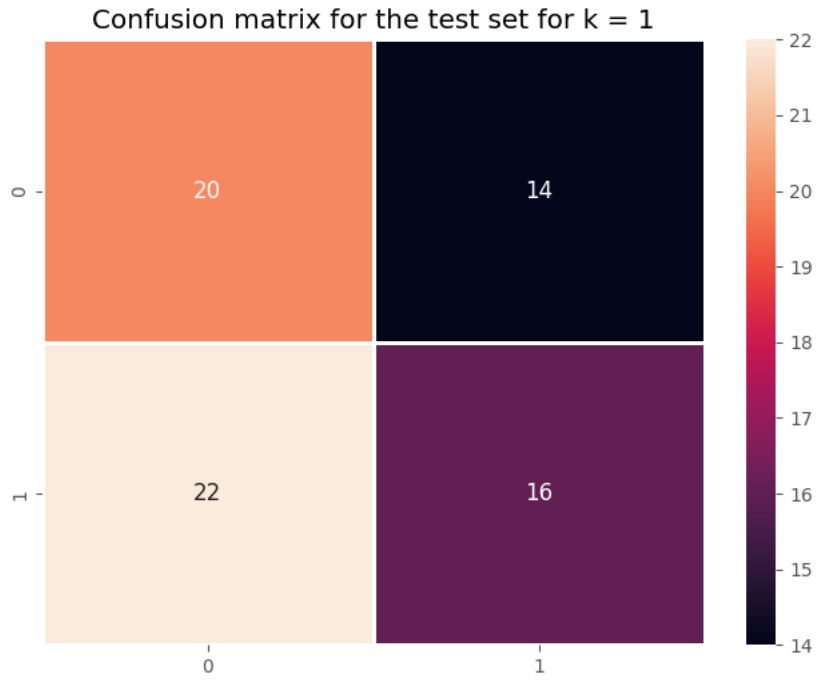

Figure 20: ROC curve and AUC score of KNN classifier for k=7

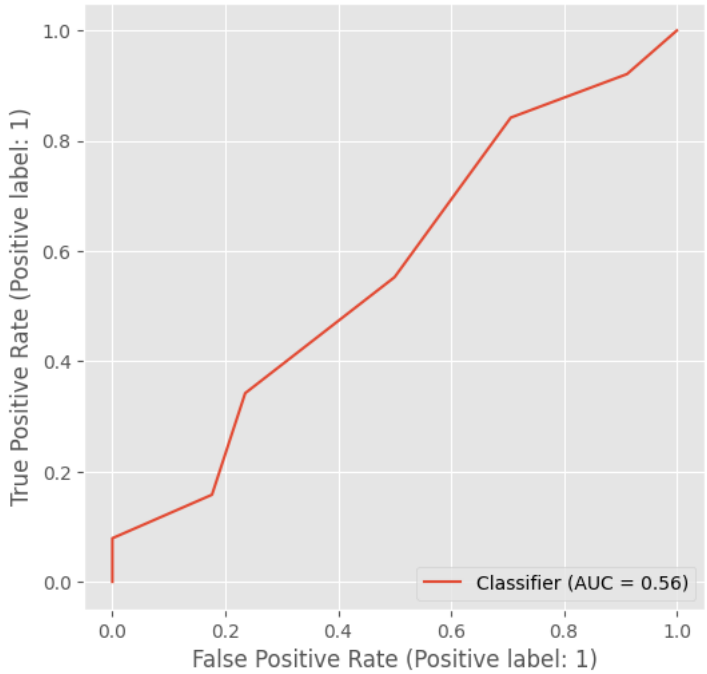

Figure 21: Confusion matrix for KNN classifier for k=7 on the test set

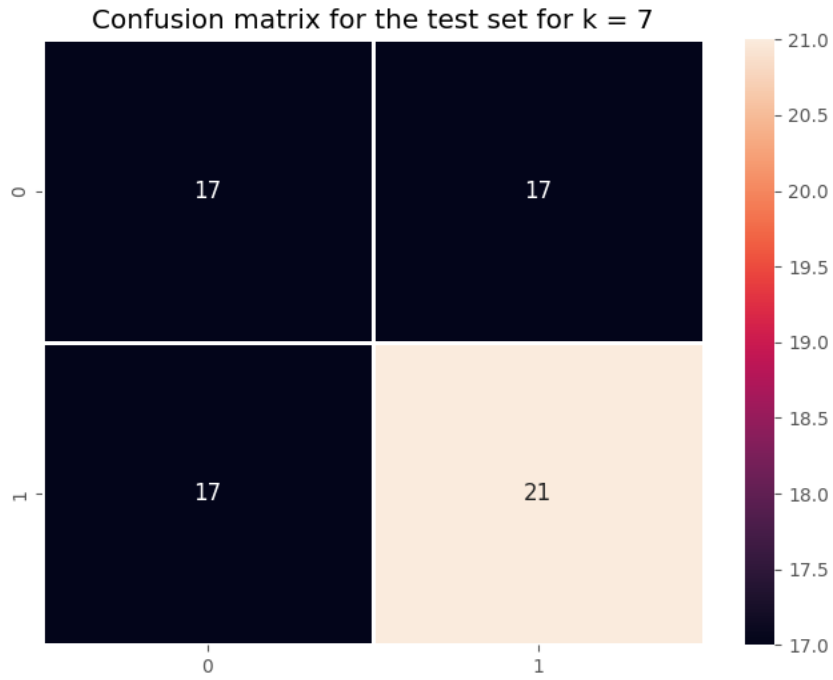

Figure 22: ROC curve and AUC score of KNN classifier for k=11

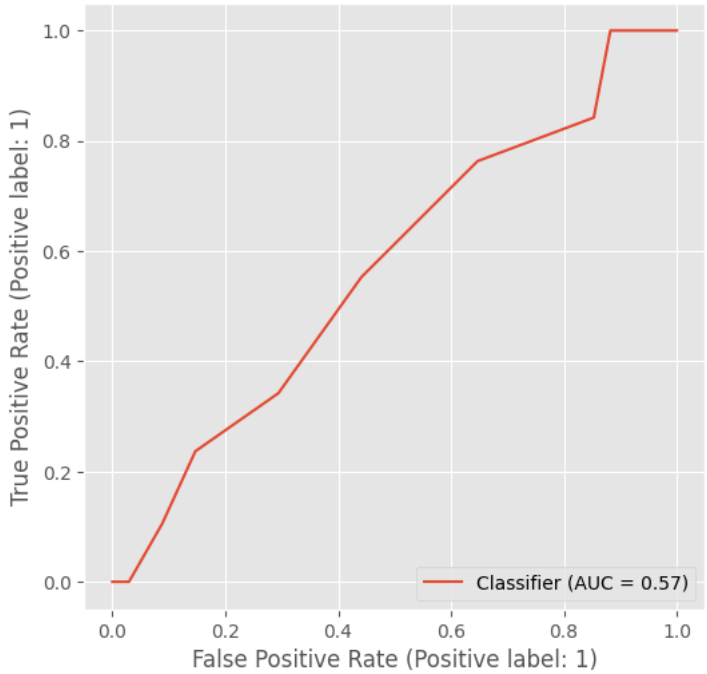

Figure 23: Confusion matrix for KNN classifier for k=11 on the test set