Summary Report for Assignment 3

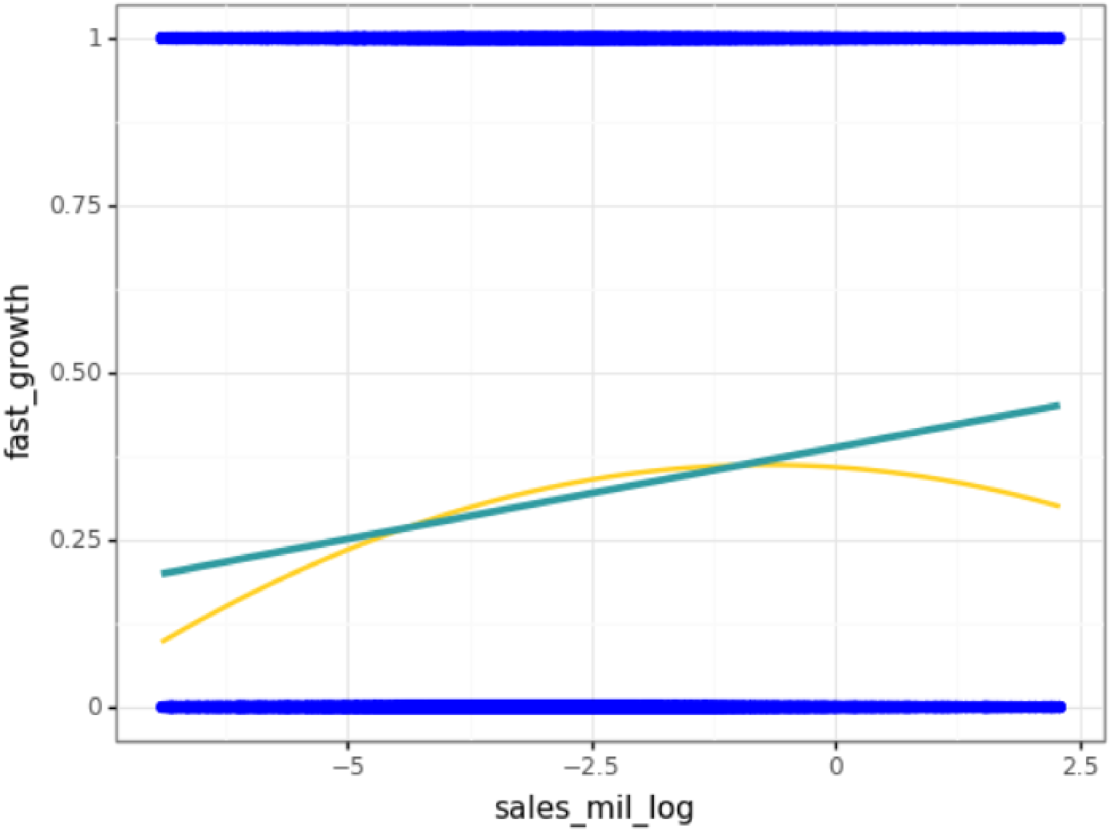
Polad Gulushov

December 6, 2021

In this assignment, we are asked to design a model to find a target growth rate and predict the firms with fast growth. Our dataset contains firm-level information in European Union. This dataset is a panel data. It contains data for the period of 2005-2016.

For this analysis, we define growth as a quarter growth in a year. As our main explanatory variable we choose sales. Let's have a look at the Figure 1.

Figure 1: Plot of fast growth against sales



Looking at this graph, one can see that there is a consistent growth of the firm as its sales increase, which is quite sensible. The growth rate approaches the half growth when firm achieves 2.5 million worth of sales.

For the prediction of growth, we chose five logit models and one random forest model. In the following Table 1, we see the results of logit model

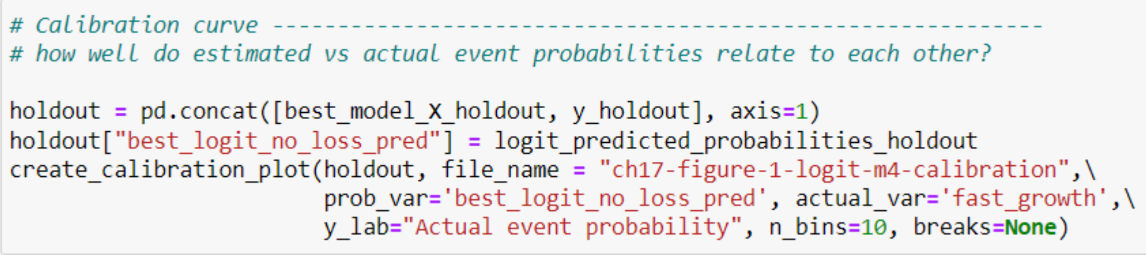
Table 1: Logit Results

	Number of Predictors	CV RMSE	CV AUC
Model1	11	0.27	0.94
Model2	18	0.27	0.95
Model3	35	0.23	0.99
Model4	78	0.26	0.98
Model5	152	0.23	0.98

As we can see from the table, when we increase number of predictors, RMSE falls and AUC increase (almost approaches one). Based on this numbers. on can choose our third model has the best predictions.

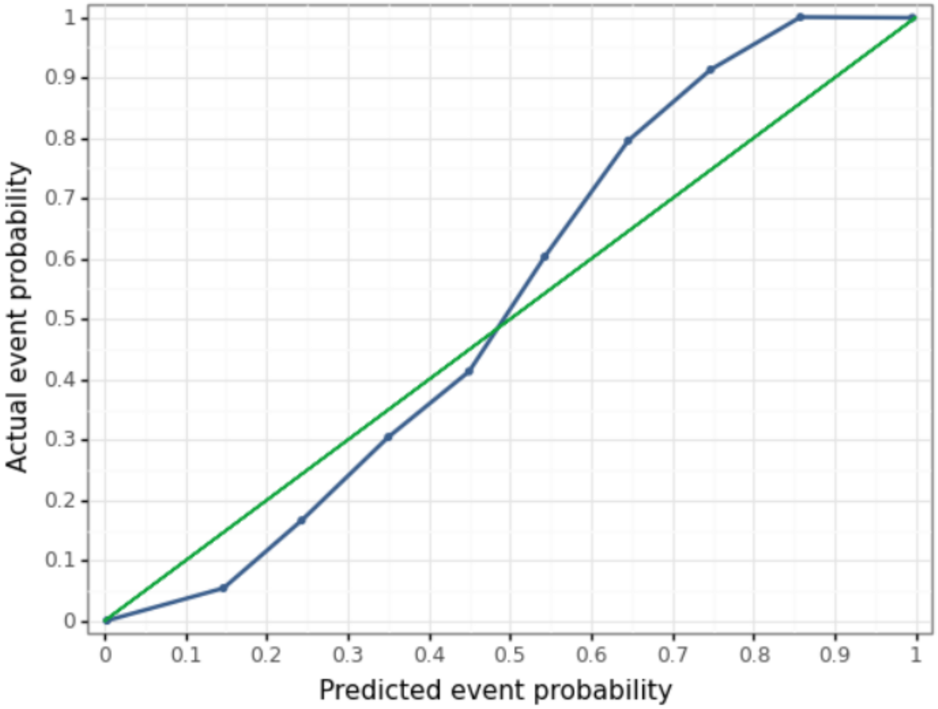
Now, let's look at the Random Forest Model. We consider two cases: First we avoid a loss function. Then we define lost function ourselves.

Figure 2: Calibration Curve



```
# Calibration curve -----------------------------------------------------------
# how well do estimated vs actual event probabilities relate to each other?

holdout = pd.concat([best_model_X_holdout, y_holdout], axis=1)
holdout["best_logit_no_loss_pred"] = logit_predicted_probabilities_holdout
create_calibration_plot(holdout, file_name = "ch17-figure-1-logit-m4-calibration",\
                        prob_var='best_logit_no_loss_pred', actual_var='fast_growth',\
                        y_lab="Actual event probability", n_bins=10, breaks=None)
```



In Figure 2, one can see the Calibration curve, which depicts how an estimated probability of fast growth is related to an actual probability of growth. Although there are deviations after the mid point, the estimated probability is mostly in line with the actual probability.

For the loss function, we assume $\frac{FN}{FP} = \frac{5}{1}$ to choose the threshold.

Finally, Table 2 presents the summary of the Cross Validation results for our models

Model	Num of predictors	CV RMSE	CV AUC	CV threshold	CV expected loss
Model1	11	0.27	0.93	0.05	0.38
Model2	18	0.27	0.95	0.09	0.26
Model3	35	0.23	0.99	0.24	0.14
Model4	78	0.26	0.98	0.28	0.14
Model5	152	0.23	0.99	0.28	0.12
Model6(random forest)	44	0.15	1	0.15	0.05

Table 2: Caption

Based on the numbers presented in Table 2 one can conclude that the random forest predicts the best.