Dimitar Argirov

KDD 1998 Cup

Content

Problem Context

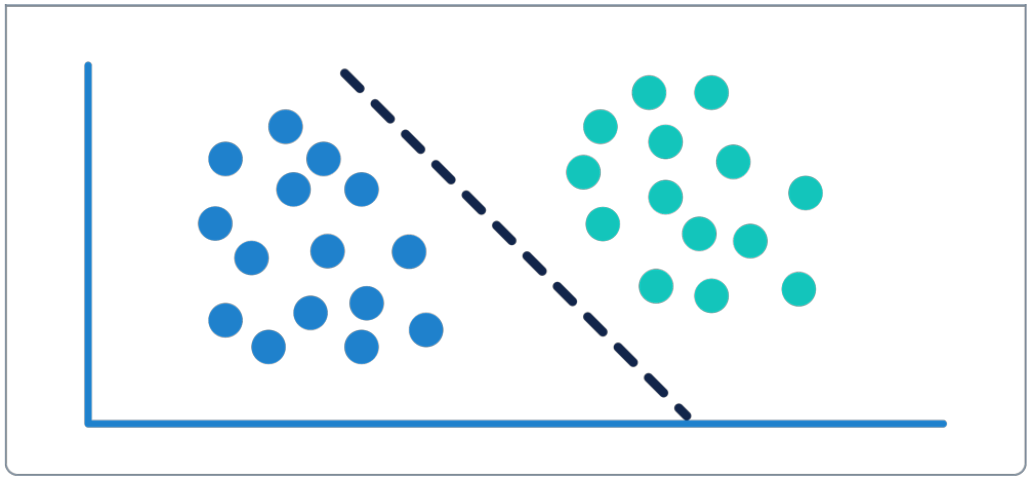
The goal of the problem is twofold: predicting the probability a person would donate and predicting the donation amount. Based on these predictions we can divise a smart way to target people such that we maximize donations.

A classification problem – use the features to predict a binary target: if a person would donate or not.
However, it could be viewed as a ranking problem:

Rank people in terms of descending probability to donate.

A regression problem – use the features to predict a continuous target: the donation amount.

Multiply the probability of donation and the predicted donation amount to get the Expected Value. Maximize the observed donation amount, given the EV and the costs of contacting people.



Data Exploration (1)

Two datasets were used:

1. training
2. validation (I treat it as test)

The datasets have 478 features and the test set is slightly larger than the training dataset (96367 and 95412 samples, respectively)

Data Quality issues:

- many variables (92) had a lot of missing values
- many variables had a mix of values corresponding to the same meaning
- some variables were in incorrect format(s)

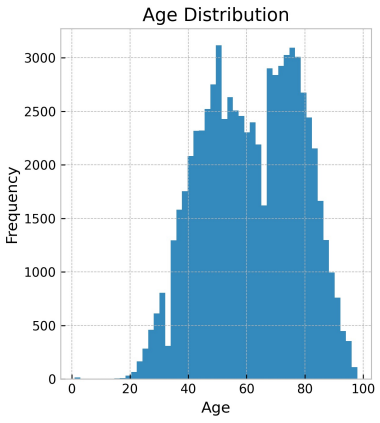
Low linear correlation between the two targets and the features: might be good to use non-linear model.

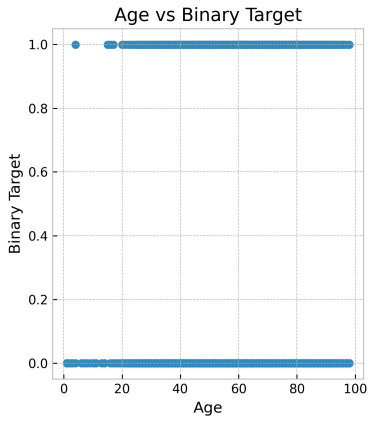
variable	number of missings	percentage of missings
AGE	23665	24.80
NUMCHLD	83026	87.02
INCOME	21286	22.31
WEALTH1	44732	46.88
MBCRAFT	52854	55.40
MBGARDEN	52854	55.40
MBBOOKS	52854	55.40
MBCOLECT	52914	55.46
MAGFAML	52854	55.40
MAGFEM	52854	55.40
MAGMALE	52854	55.40
PUBGARDN	52854	55.40
PUBCULIN	52854	55.40
…	…	…

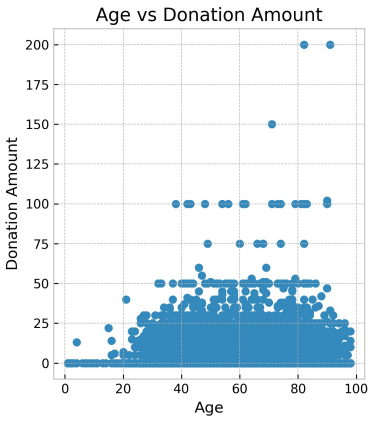
Data Exploration (2)

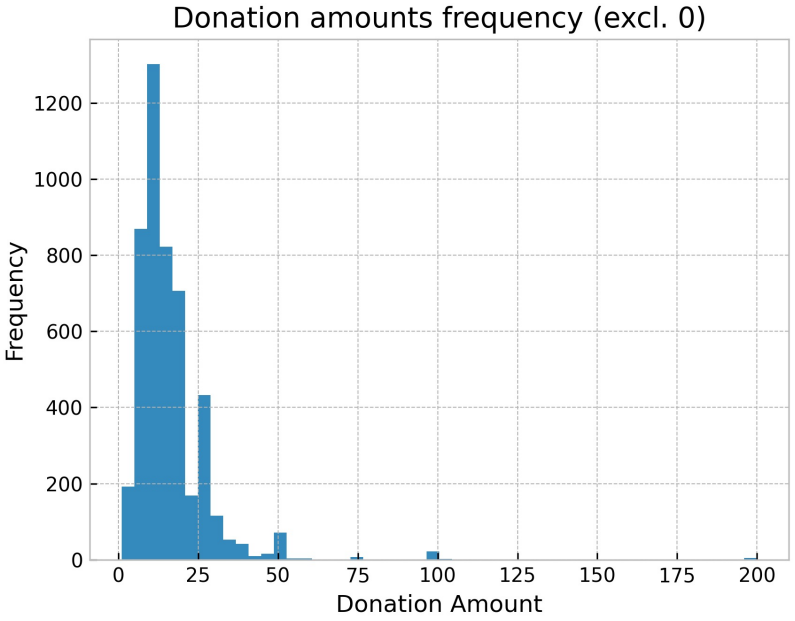
There are some interesting aspects of the merged dataset:

- The target variable distribution is heavily imbalanced
  for the binary target (is donor Y/N): 95% : 5%
- Most donations are for relatively small amounts
- Many donations in the test set are 0
- The age variable seems to correlate non-randomly with donors and donation amounts









Key Data Takeaways

1. The data is imbalanced
   - Classificcation: Accuracy is not a relevant metric (AUC?, F1?)
   - Classificcation: Might want to do over/under sampling or use weights
   - Regression: Training should be done only on the donors subset

2. The main objective from a business standpoint is:
   - Maximize donation amount, given engagement costs

3. A non-linear modelling approach might be preferred (trees, neural network, etc.)

4. When processing the data, we should not take information from the test set (encoding, feature engineering, etc.)

5. The end product of the model in production should aid managers in the charity: figure out which people they should contact based on expected donation and contact costs

Data Preprocessing

- The dataset contains Date varialbes in the format YYMM (string columns). These can be converted to date format.
- For some date variables (LASTDATE, MAXADATE, etc.) a difference with the present in terms of days can be calculated and used as a feature. The present is assumed to be 1998/07 – the month of the KDD cup in 1998.
- Many string variables are actually categories and they should be treated as such – cast to categorical and encode on the train set; apply the encodings to the test set
- Recode some categorical variables (say ' ' and 'X' to 0 and 1)
- Drop samples that have invalid MAILCODE – we cannot contact these people even if they have a high expected donation
- Extract information from multi-bit string variables (SOLP3, RFA, DOMAIN, MDMAUD)
- Fill cases of unknown GENDER based on the title variable (TCODE)
- Fill cases of missing AGE with age calculated based on the date of birth (DOB)
- Group small categories under 'other' for some categorical variables
- Fix the formatting of some variables like ZIP by standartizing the format
- Drop some variables that are clearly redundant or that have been used to extract features already
- Drop ALL variables related to the respondent's regional census (controversial)

- When using regressions (logit, OLS, etc) the missing values need to be filled and the variables need to be standartized to guarantee accurate and robust results. However, tree based methods can deal with missings and outlier values.
- I have chosen not to apply transformations and not to fill missing values because of that. These techniques can also bias the training data and are sometimes difficult to properly apply in production

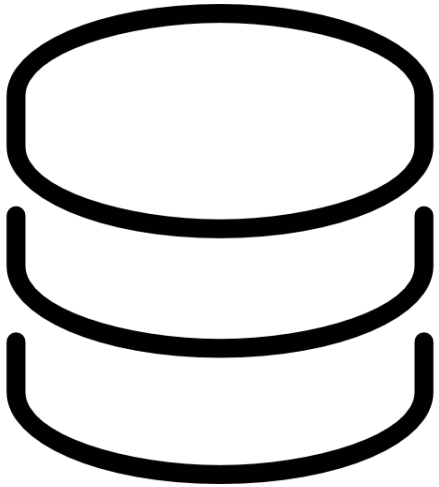
Training and Testing Datasets

When we split the data for the classification task, we want to make sure that each subset is stratified with respect to the target. If we do cross validation this is also the case for each fold.

We can apply a similar logic to the regression task by binning the continuous target and stratifying the split with respect the average value of each bin.

We also need to make sure that we are not training on information we wouldn't see in a production setting to predict the target. Thus, data manipulation should be done separately on the training and the test set.

Solution:
Use the train set for training and do cross validation within it. Use the test set for testing final model performance.



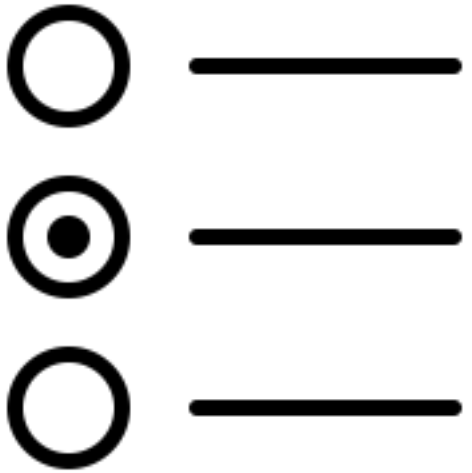
Model & Feature Selection

We should select features for the two tasks (classification and regression) separately.
When selecting features we can consider feature correlation, feature variance, univariate performance, business sense.

For classification:
I have used cross validated recursive feature elimination based on a lightGBM classifier. All model parameters are left to default except class weights and importance type.

For regression:
I have used cross validated recursive feature elimination based on a lightGBM regressor.

- lightGBM is considered state of the art for tabular problems
- RFE is more precise then crude approaches (correlation, variance based approaches), albeit computationally more expensive

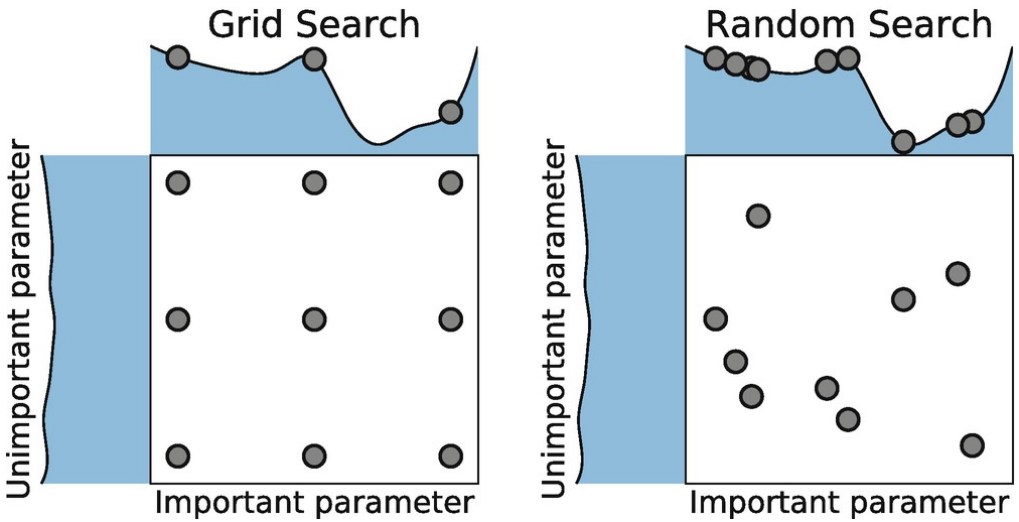


Based on the two RFEs, features are selected for both tasks. These features and used in hyperparameter tuning and in the respective final models.

Hyperparameter Tuning

For each of the two models (classifier and regressor) I have ran a randomized grid search to find optimal parameters. This ensures (somewhat) optimal model performance and minimizes overfitting.

Other alternatives to randomized grid search are full or Bayesian grid searches. These can sometimes provide better results, but are computationally more expensive.

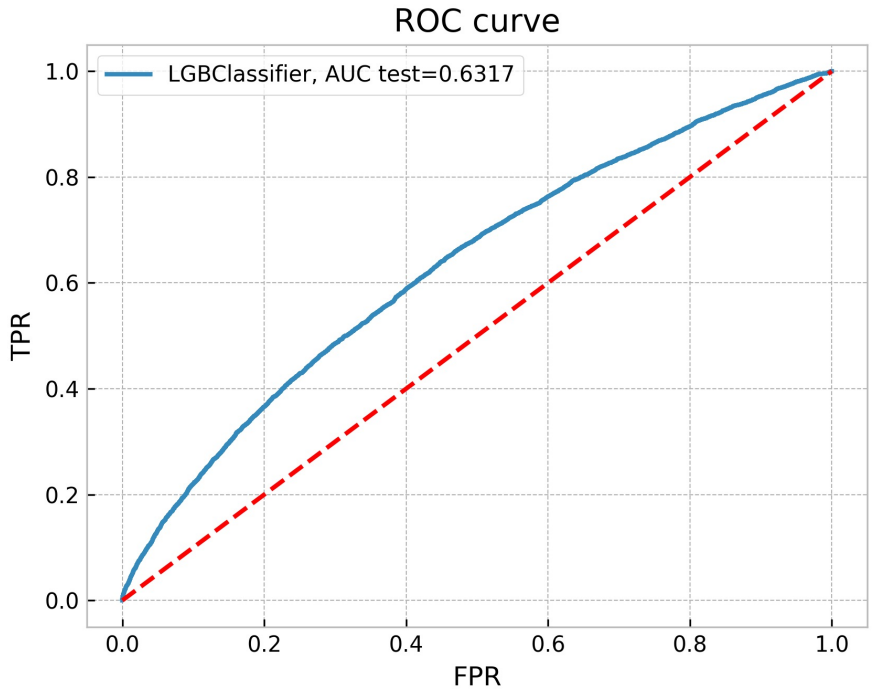


Final Model: Classification

Based on the optimal parameters found during the cross validated randomized grid search and the proceeding feature selection, a final model was trained on the entire training set.

This model was then used to predict probabilities of donating for the test set.

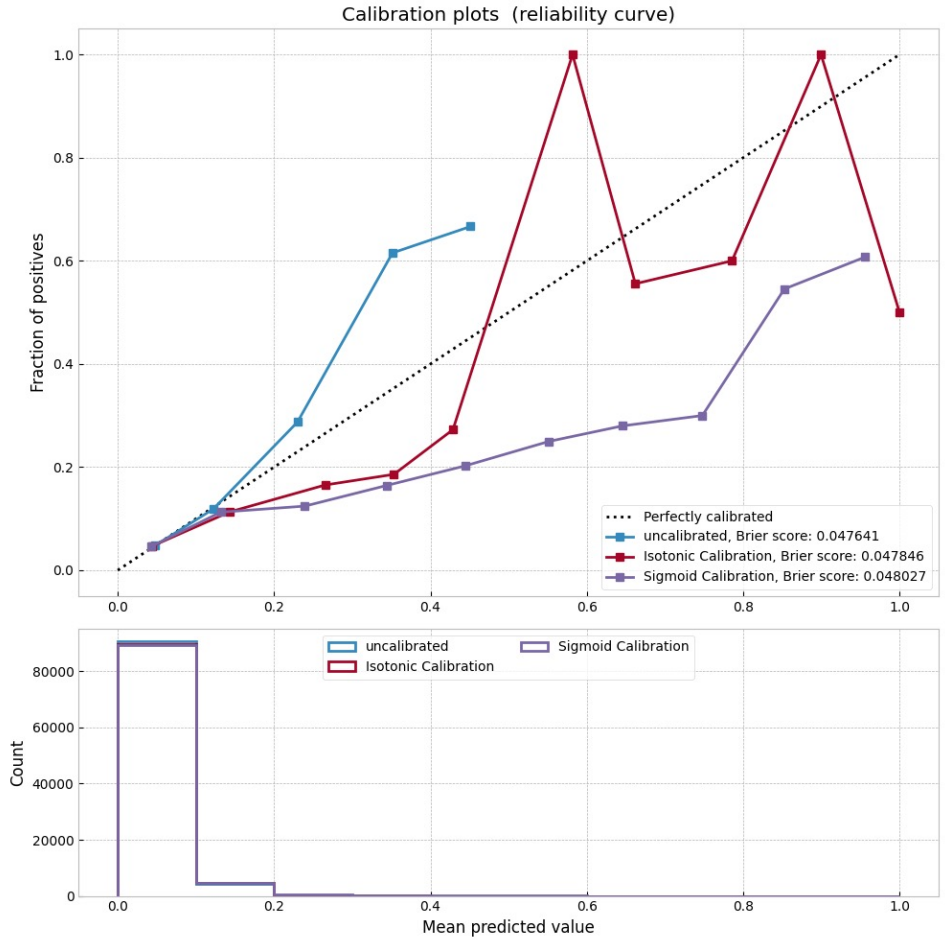
The performance was measured in terms of AUC.



Classifier Calibration

Given the nature of the prediction task a well-calibrated classifier is important. Therefore, model calibration was attempted via isotonic and sigmoid calibrations. The resulting classifiers were evaluated in terms of Brier score.

Nevertheless, the raw predictions of the LGBMClassifier turned out to have the lowest Brier score. Thus, they were kept as the best probability predictions.

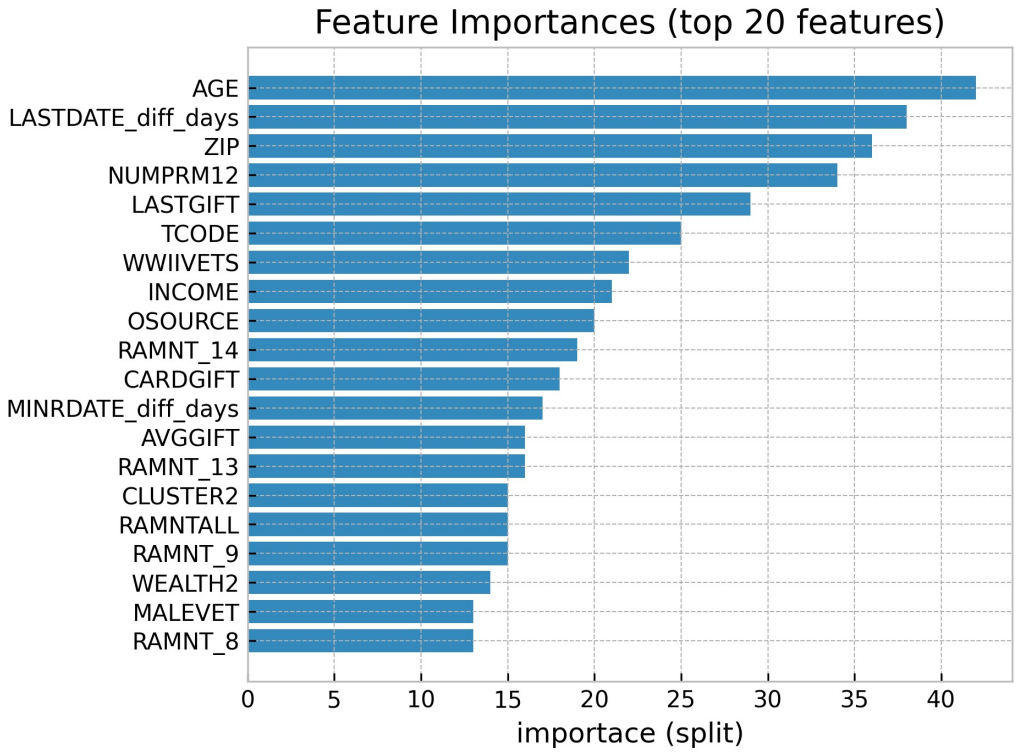


Classifier Possible Improvements

Perhaps, additional work on the features used for the classification could improve the AUC and the calibration of the model.

It is worth exploring particularly the following:

- A larger parameter grid
- Binning some categorical variables (i.e AGE) and iteratively trying out how different transformations affect the model

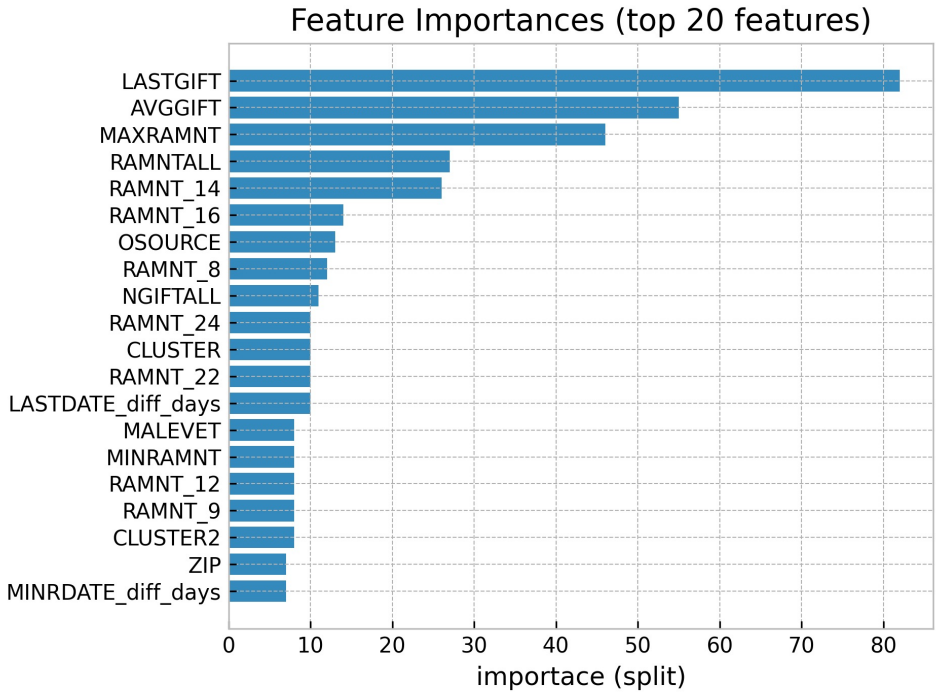


Final Model: Regression

Based on the optimal parameters found during the cross validated randomized grid search and the proceeding feature selection, a final model was trained on donors portion of the training set. (4791 samples)

This model was then used to predict donation amounts in the test set.

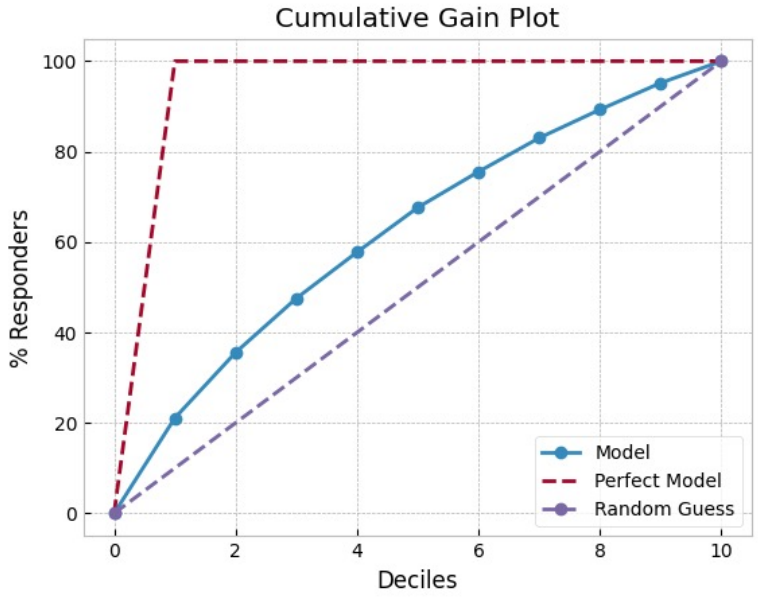
The performance was measured in terms of MSE.

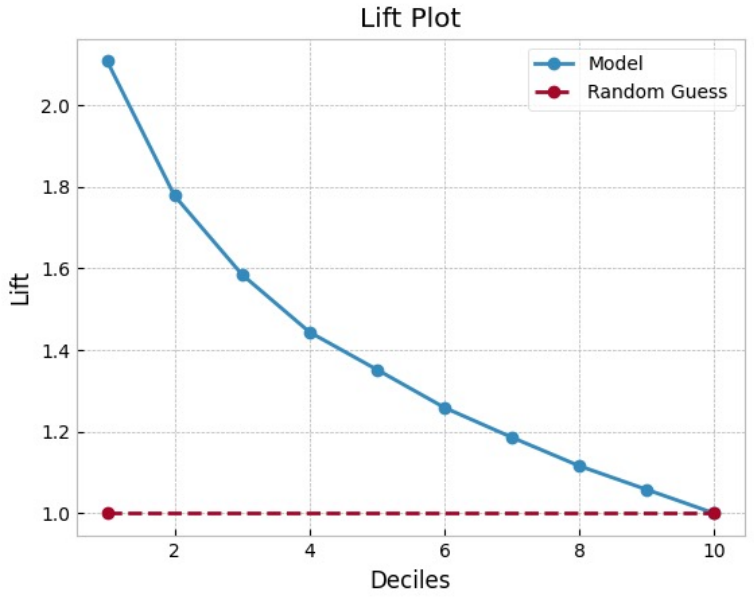


Solution Evalutation: Classifier

From a business standpoint the classifier can be evaluated by cumulative gain and lift:

- What percentage of donors will we engage if we engage the top N% of samples according to their predicted probability of donating (highest first)





Solution Evalutation

From a business standpoint the complete model (classifier and regressor) can be evaluated in terms of actual gain from expected value (EV):

Sum (the actual donation amount – cost of donating) over all records for which the expected revenue is over cost of donating.

Mail Cost	Total EV	Max. Donation	Min. Donation	Std. Donation	# Engagements	% Engagements
1	7091.95	499	-1	8.34	17599	18.54
5	1022	195	-5	29.97	149	0.16
0.68	16676.07	499.32	-0.68	5.79	51865	54.64

TABLE 2: KDD-CUP-98 Summary of Evaluation Results: Total Profits for Records with Predicted Donation > $0.68

Participant	N*	MIN	MEAN	STD	MAX	SUM**
GainSmarts	56,330	-$0.68	$0.26	$5.57	$499.32	$14,712
SAS	55,838	-$0.68	$0.26	$5.64	$499.32	$14,662
Quadstone	57,836	-$0.68	$0.24	$5.66	$499.32	$13,954
CARRL	55,650	-$0.68	$0.25	$5.61	$499.32	$13,825
Amdocs	51,906	-$0.68	$0.27	$5.69	$499.32	$13,794

Possible Improvements

More hyperparameters could be used in the optimization. Furthermore, instead of using randomized grid search, we could use Bayesian optimization on continuous intervals.

Iteratively preprocess the data in different manners and see how that affects the final outcome.

Try a different methodology. Sequential neural networks have been shown to give good results when it comes to imbalanced datasets.

Build a baseline model to measure performance against (logit and OLS).