Springboard

Springleaf customer adoption

Capstone 2 Project Summary

Robert Ciesielski

Mentor: Dipanjan Sarkar

May 2021

Introduction

- In 2015, Springleaf, a company that provides personal loans and optional insurance products, released to the Kaggle community an anonymized dataset with about 2,000 features describing 145,000 customers.
- Kaggle challenge: predict which customers will respond to a direct mail offer.
- Our goals:
  - Find an optimal method to reduce the dimensionality of this wide dataset.
  - Compare the performance of various ensemble-based models, such as Random Forest, two scikit-learn Gradient Boosting models, XGBoost and LightGBM.
  - Model customer response also using a feed-forward neural network architecture.

Dataset

Provided as a csv table.

Classification problem to solve.

Customer response	N. of rows	
No (0)	111,458	(76.7%)
Yes (1)	33,773	(23.3%)

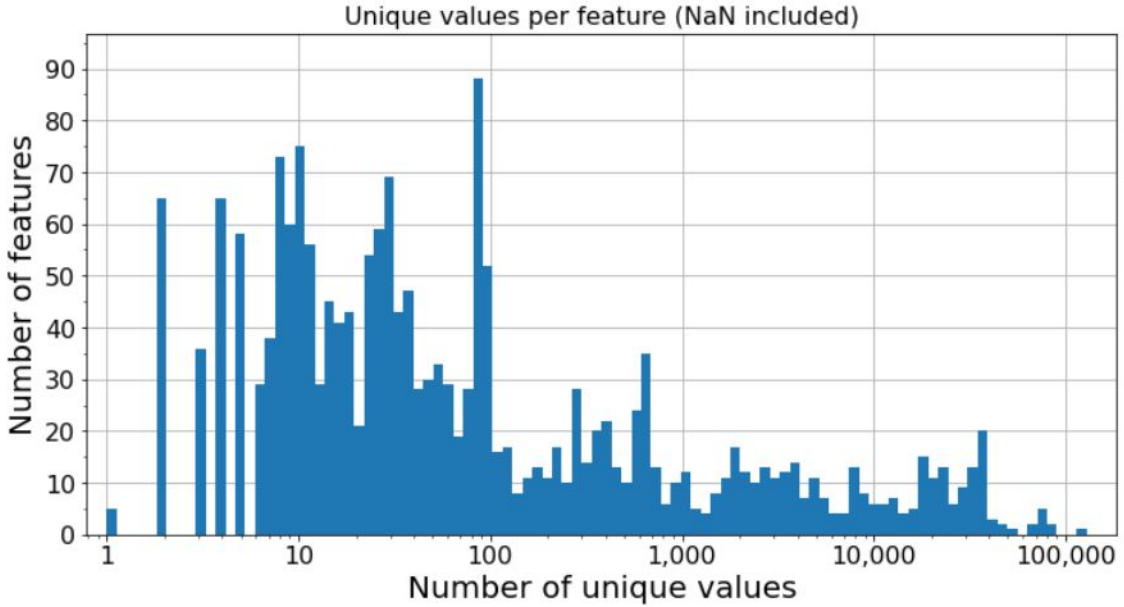
Slightly unbalanced data.

Feature type		N. of features
Numerical:	Integers	1,837
	Floats	10
	DateTimes	18
Categorical:	Strings	17

Dominated by Integer-type features.

Data Wrangling

Feature unique values



- Start with inspecting features with small number of unique values.
- Remove those with one unique value (no variance).
- Features with two unique values:
  - If they contain a NaN and another number -> replace by them 0 and 1.
  - Replace -99999/0 by 1/ 0.
  - Convert False/True to 0/1, etc.
- After the conversion, look for duplicates -> 51 features removed.

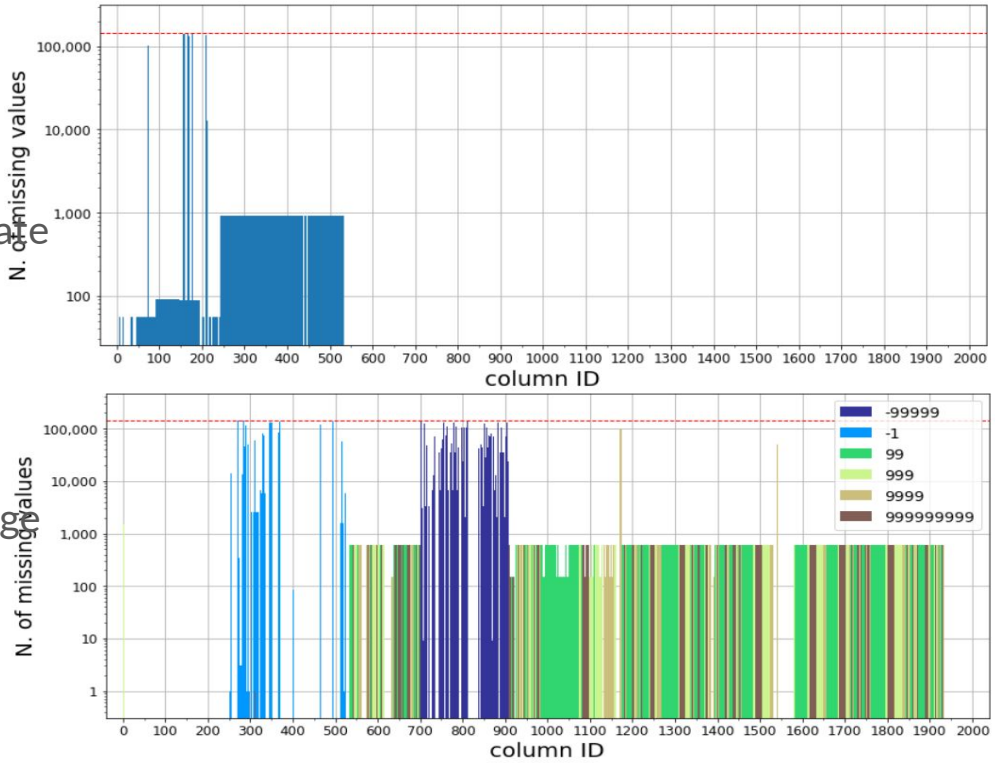
Data Wrangling (2)

Missing values (MV)

- Only 530 columns contain MV explicitly marked as NaNs; for most of them the NaN rate is < 1 percent.

  -> Replace NaNs with -1

- MV are also encoded as -1 or a very small/large integer number.



Exploratory Data Analysis & Feature Engineering

- A lot of work, mostly for 17 categorical and 18 DateTime features :-)
- E.g. converted a feature with 12,000 unique city names into two new features with the city population size and density, using an additional dataset from the US Geological Survey and US Census Bureau, etc.
- All features are documented in the report.

- This anonymized dataset contains healthcare data related to the hospital treatment of a cancer disease. The patients reside in most of the US, except the states in New England.

Dimensionality Reduction

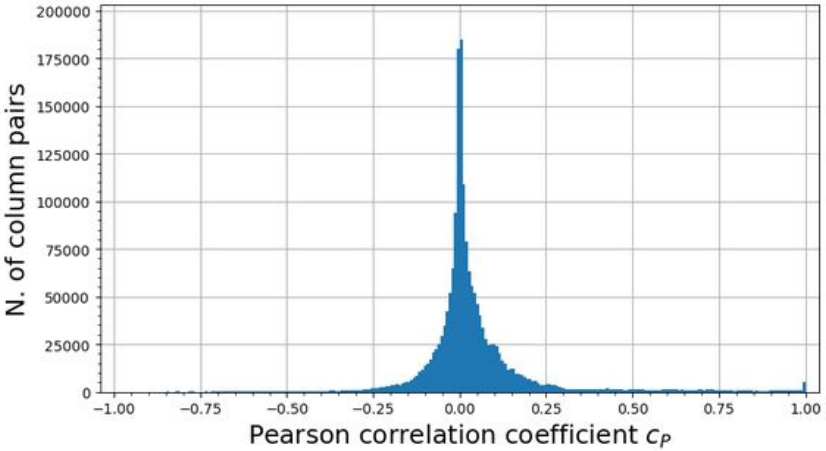
Four methods to reduce the dimensionality of the dataset were explored:

1. Feature correlation matrix
2. Principal Component Analysis
3. Logistic Regression with L1 Regularization
4. Recursive Feature Elimination

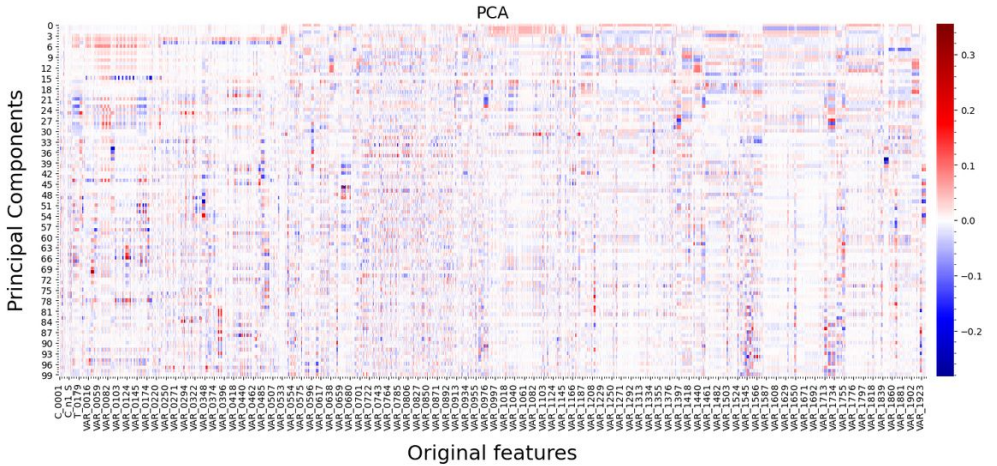
Only method 3 and 4 worked.

Dimensionality Reduction (1)(2)

Feature correlation matrix

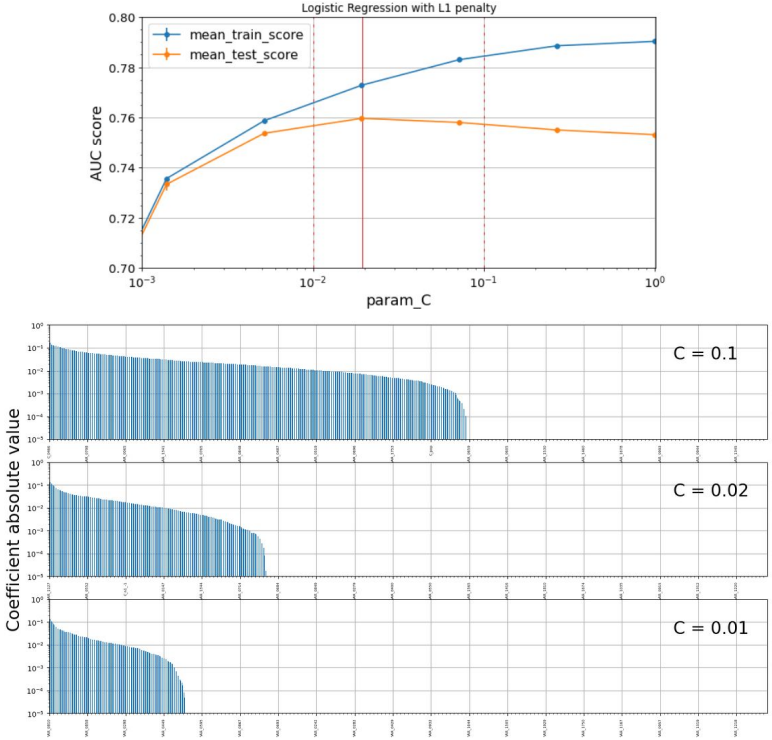


Principal Component Analysis

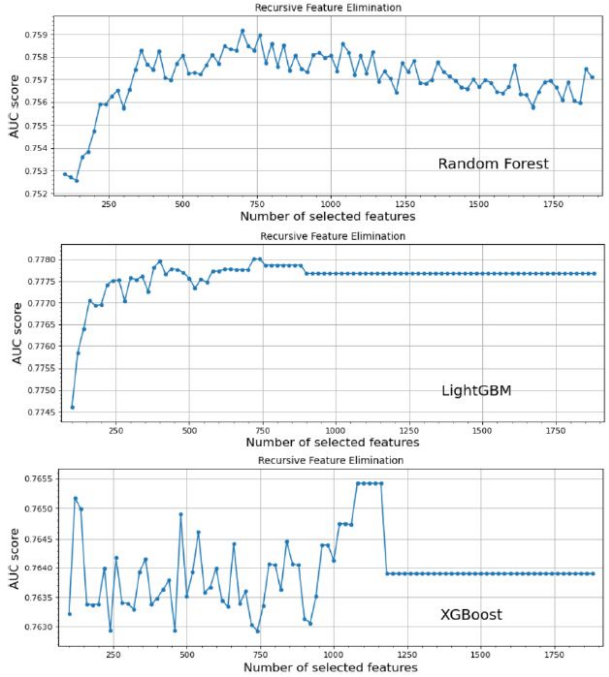


Dimensionality Reduction (3)(4)

Logistic Regression with L1 reg. (LR,L1)



Recursive Feature Elimination (RFE)



Dimensionality Reduction (5)

	Method	N. of columns
•	(none)	1880
	LR, L1-0.1	1096
•	LR, L1-0.02	569
•	LR, L1-0.01	359
	RFE, XGB	1080
•	RFE, RF	700
•	RFE, LGBM	360

Managed to reduced the dataset from 1880 to 360 features.

Dot indicate samples used for modeling.

Modeling

Classification problem. Performance metrics - area under the ROC curve (AUC).

1) Logistic Regression (C=0.02): AUC = 0.767 (baseline)
2) Random Forest (out of the box): AUC = 0.753
3) Gradient Boosting models (out of the box):

GB - sckit-learn GradientBoosting
XGB - XGBoost
XGBh - XGBoost+histogrammed features
LGBM - LightGBM
HGB - sckit-learn HistGradientBoosting

	auc_roc	acc.	prec.	recall	time (min.)
GB	0.777	0.794	0.628	0.281	18.2
XGB	0.777	0.795	0.633	0.279	1.9
XGBh	0.778	0.794	0.633	0.276	0.9
LGBM	0.778	0.795	0.629	0.287	0.4
HGB	0.777	0.795	0.629	0.285	1.0

Compared performance of various out-of-the-box GB models.
LGBM and XGBh seem the best and the fastest, and will be used further.

Hyperparameter tuning

3 sets of model parameters are tested:

- Default (D) - out-of-the-box sets.
- Manual Tune (MT) - one-dimensional parameter scan with GridSearchCV.
- Bayesian Optimization (BO) - multi-dimensional optimization.

Full dataset

Hyperparameter set	RF score | ratio	XGBh score | ratio	LGBM score | ratio
Default (D)	0.753 | 1.00	0.7779 | 1.00	0.7783 | 1.00
Manual Tune (MT)	0.779 | 1.03	0.7868 | 1.01	0.7884 | 1.01
Bayesian Opt. (BO)		0.7969 | 1.02	0.7985 | 1.03
D	0.97	1.00	1.001
MT, ratio to XGBh	0.99	1.00	1.002
BO		1.00	1.002

ratio to D

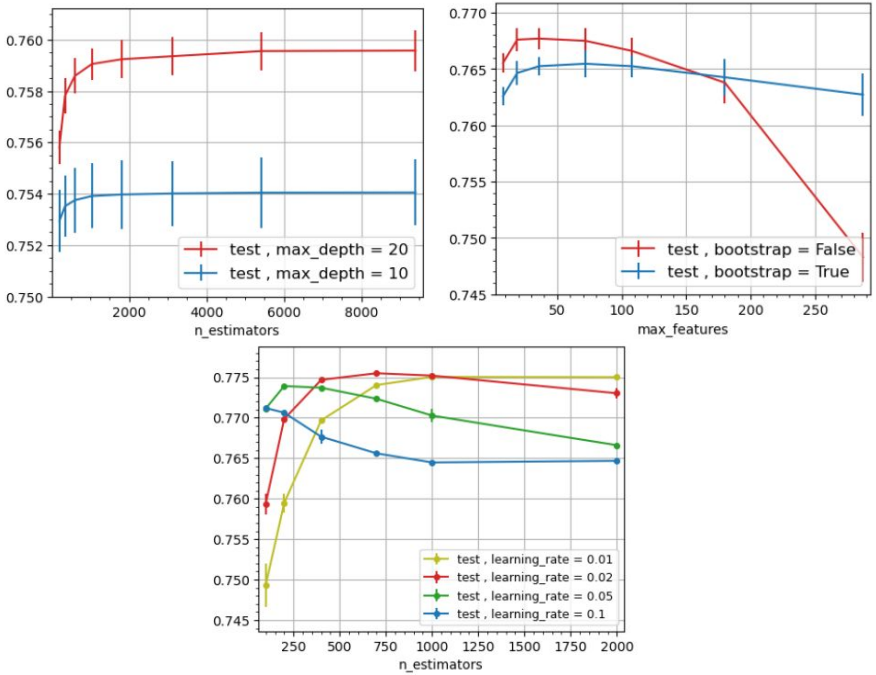
ratio to XGBh

- For GB model a relatively quick Manual Tune improves AUC score by 1%.
- The time consuming Bayesian Opt. adds additional 1-2%.
- RF has the worse performance, while LGBM systematically outperforms XGBh.

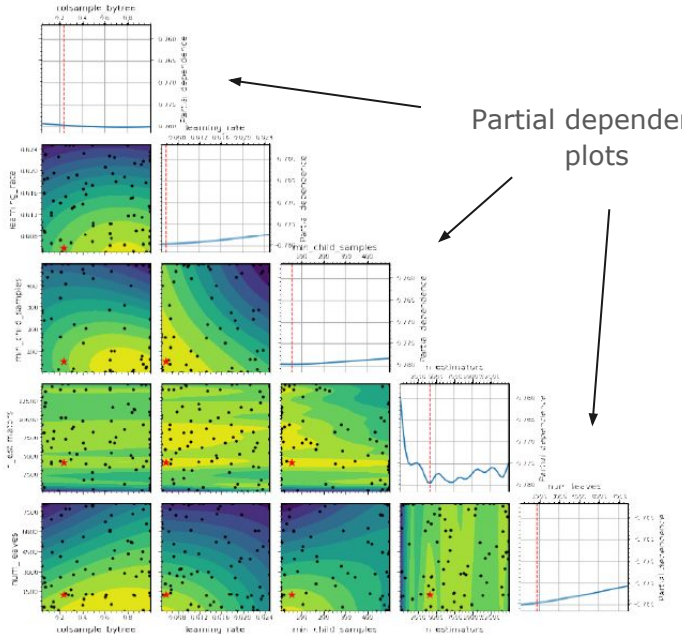
Hyperparameter tuning (2)

Examples

Manual Tune (MT)



Bayesian Optimization (BO)



Impact of dimensionality reduction on model performance

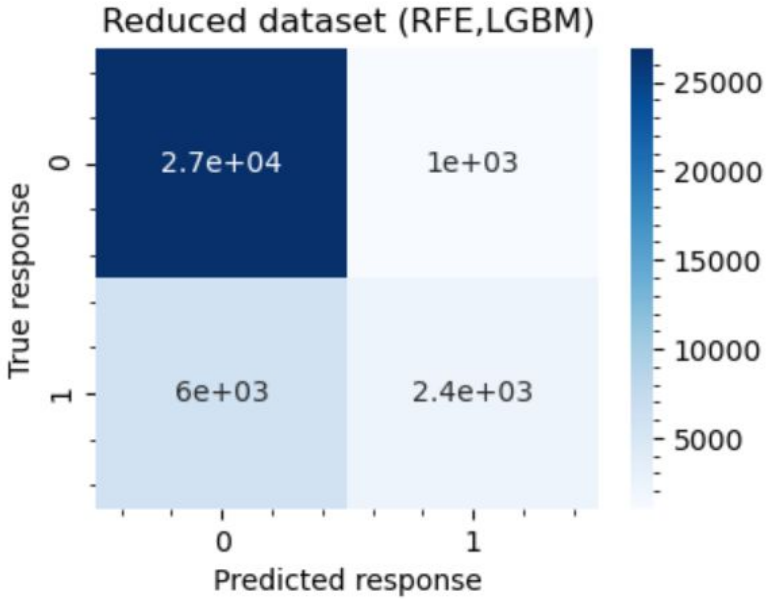
Use LGBM model and Bayesian Opt. to check how the score changes with reduced data samples.

Dataset	(n. cols)	auc_roc	ratio to Full
Full	(1880)	0.7985	1.000
LR, L1-0.02	(569)	0.7962	0.997
LR, L1-0.01	(359)	0.7959	0.997
RFE, RF	(700)	0.7971	0.998
RFE, LGBM	(360)	0.7998	1.002

Reduced dataset with 360 features, selected using the RFE method with LGBM model, gives 2-permille better result than the full dataset and is used to predict customer response.

Final prediction

Confusion matrix

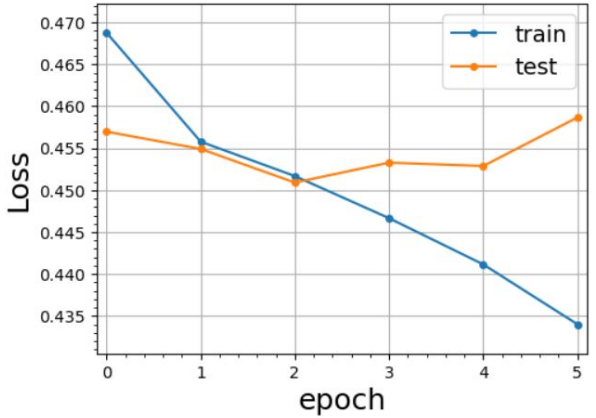


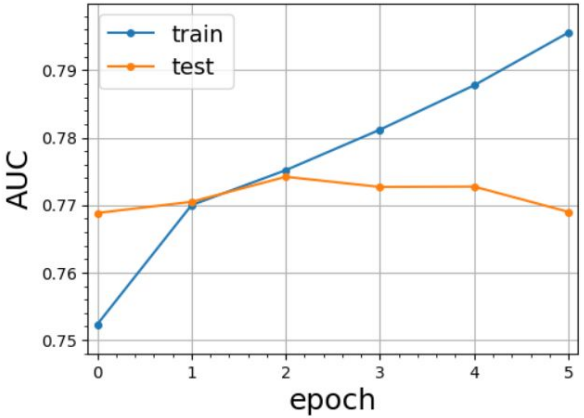
The best model allows to target 30% of the customers of interest (recall) and expect positive responses for 70% of the offers sent to them (precision).

Modeling, Neural-Network-based classifier

3 or 4 dense layers with various number of parameters, and the binary classifier in the last layer.

Reduced and full dataset.





	Model	auc_roc
0	NN3_360_360_360_1	0.7738
1	NN3_360_180_90_1	0.7742
2	NN3_360_720_180_1	0.7737
3	NN3_360_90_45_1	0.7728
4	NN4_360_180_90_45_1	0.7740
5	NN3_1800_940_470_1	0.7704

The best model (1) performs better than LR model, but has worse performance than RF and GB models.

Summary and Future Work

- The aim of this study was to predict customer adoption using binary classification methods and a wide dataset with 2,000 anonymized features.
- We explored the performance of several decision-tree based ensemble models, such as Random Forest, XGBoost, and LightGBM, as well as Logistic Regression and neural-network-based models.
- LightGBM outperforms other models for this dataset.
- The best score (area-under-the-ROC-curve = 0.8) is obtained after reducing the dimensionality of the dataset to 360 features using the Recursive Feature Elimination technique.
- The best model allows the Sprigleaf company to target 30% of the customers of interest (recall) and expect positive responses for 70% of the offers sent to them (precision).

- Future work should focus on improving the relatively low recall value, with additional domain knowledge provided by Springleaf by unveiling the data features.