Predicting Credit Card Default with scikit-learn

By Gideon Blinick

Overview

- Problem:
  - Using information about a client including demographic information (age, sex, marital status) and financial information (credit limit, repayment statuses, monthly bills, and monthly repayments), are we able to build models that can predict if new customers will default?
  - Will our new models perform better than simply predicting everyone won’t default, since that happens the majority of the time?
  - Which model will perform best?

Overview (continued)

- Client: Credit-card companies or any credit-lending institution more generally.
- Motivation: Credit-card companies lose money every time they extend credit to a customer who does not pay back (i.e. defaults). If credit-card companies can figure out which customers will not pay back, they can save a lot of money.

Data-Science problem and Dataset Description

- Formulation of the problem: This is a supervised-learning, classification problem. The labels to predict are 0 (no-default), and 1 (default).
- Dataset:
  - On Kaggle: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home
  - And UCI Repository: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
  - Paper about the dataset that provided motivation for the project: https://bradzzz.gitbooks.io/ga-dsi-seattle/content/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/DefaultCreditCardClients_yeh_2009.pdf:

Dataset Description (continued)

- 23 features:
  - LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
  - SEX: Gender (1=male, 2=female)
  - EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
  - MARRIAGE: Marital status (1=married, 2=single, 3=others)
  - AGE: Age in years
  - PAY_X: 6 variables (X goes from 1 to 6) representing how many months behind on payment a customer is in month X.
  - BILL_AMTX: 6 variables (X goes from 1 to 6) representing the amount of bill statement in month X.
  - PAY_AMTX: 6 variables (X goes from 1 to 6) representing the amount of previous payment in month X.
- 1 target variable:
  - default.payment.next.month: Default payment (1=yes, 0=no)

Data Wrangling

- First, we checked for Null or missing values. Everything fine.
- Second, we checked for outliers. We did find statistical outliers (greater than Q3 + 1.5*(IQ range) or less than Q1 – 1.5*(IQ range)), but they made sense and did not look like errors.
- Third, we did find values that needed to be recoded, because they either violated the dataset description, or did not make sense. (continued on next slide)

Data Wrangling (continued)

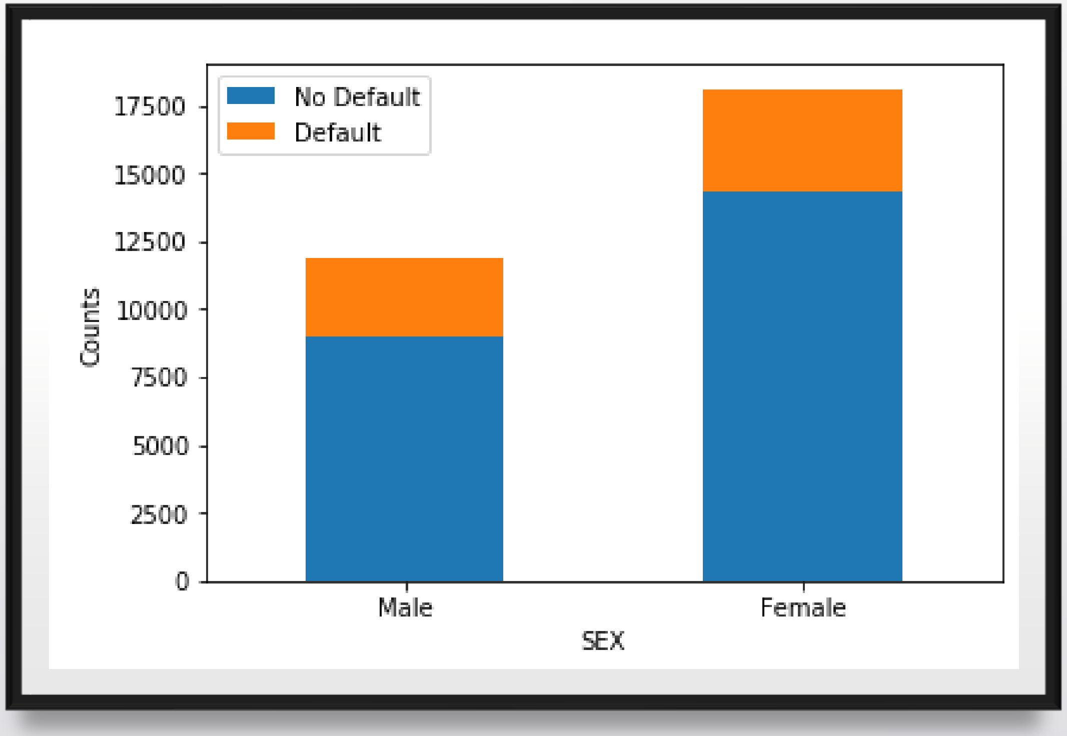
- Re-coded values:
    - PAY_X (repayment status) columns: contained values of-2, -1, and 0. By the description, should have only contained values of -1. However, all these values were set to 0 to represent 0 months behind payment (i.e. on time).
    - MARRIAGE column: Contained a few values of 0 (not part of dataset description). Recoded to 3, representing “other” for marital status.
    - EDUCATION column: Contained values of 0 (not part of dataset description). Furthermore, it contained 2 different codings for “unknown” and 1 coding for “other”. The choice was made to recode all these values to “other”.

Exploratory Analysis and Findings

- The first thing done in the Exploratory Data Analysis phase was to examine correlations between our features and target variable.
- We discovered that default is only somewhat correlated (an absolute correlation greater value than an 0.1) with 2 variables: LIMIT_BAL (the amount of credit a customer has) and the PAY_X variables.
  - Default was negatively correlated with credit (-0.15), meaning that individuals with more credit defaulted less. A reason for this could be that the credit-issuing agency would only extend credit to individuals they knew were less likely to default
  - Default was positively correlated with repayment status, meaning individuals who were more behind on payments were more likely to default. Furthermore the correlation increased with time, meaning repayment status at later times mattered more for predicting default than at earlier times. The correlations ranged from 0.24 in the earliest month to 0.4 in the latest.

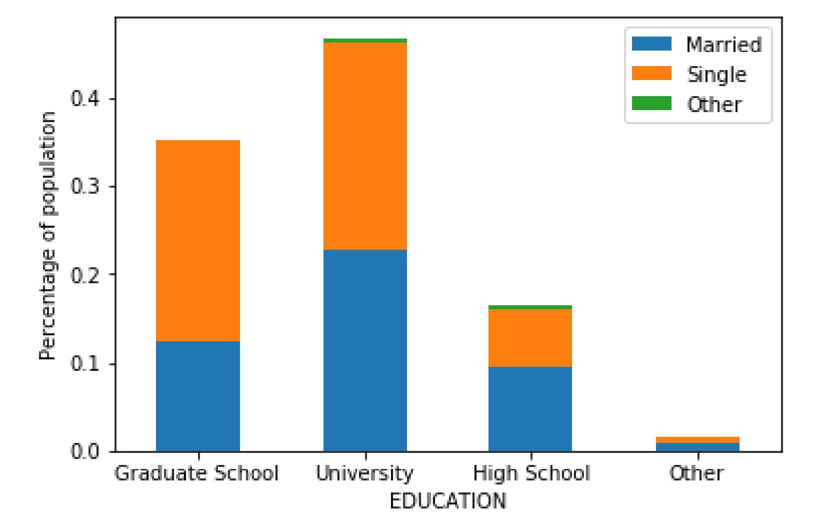
Exploratory Analysis and Finding (continued)

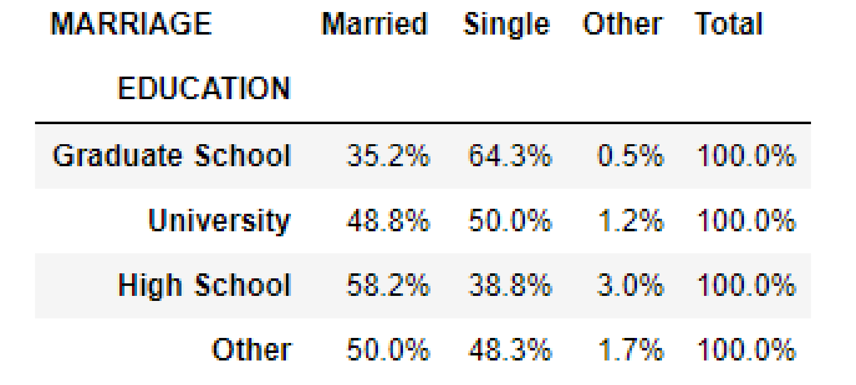
- In the second part of our exploratory stage, we examined connections among variables, not all of which will be covered in this presentation.
- The first notable finding was that men default at a higher rate than women.



Exploratory Analysis and Finding (continued)

- We also found that people with only a high school education have higher rates of marriage than people with graduate school education.

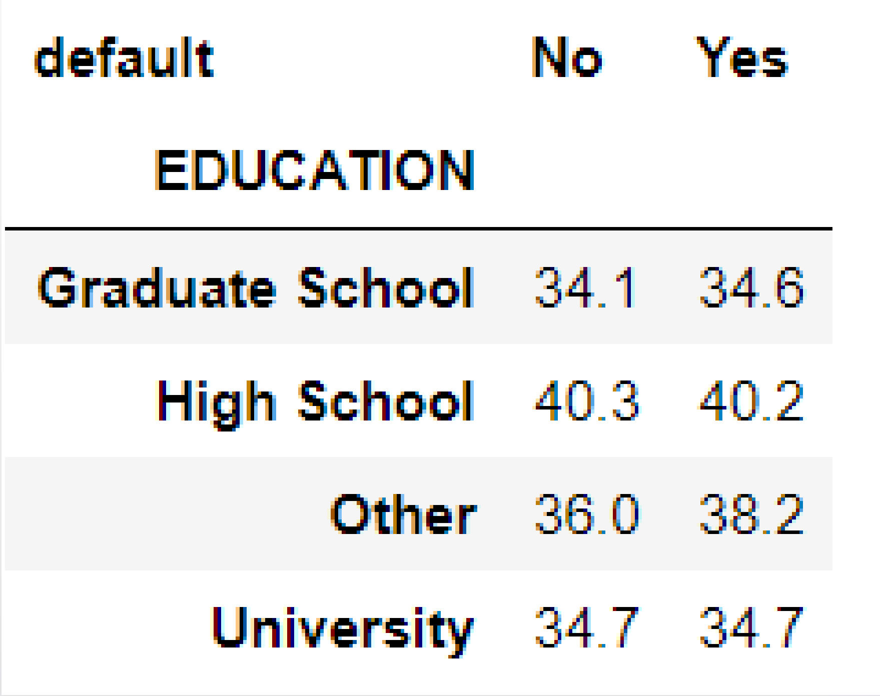




MARRIAGE	Married	Single	Other	Total
EDUCATION				
Graduate School	35.2%	64.3%	0.5%	100.0%
University	48.8%	50.0%	1.2%	100.0%
High School	58.2%	38.8%	3.0%	100.0%
Other	50.0%	48.3%	1.7%	100.0%

Exploratory Analysis and Finding (continued)

- A third finding was that while age is certainly related to education with more educated individuals being younger in our dataset, it was not a significant factor in distinguishing between defaulters and non-defaulters.
- The table shows average age for individuals, broken down by education and whether they defaulted or not.



default	No	Yes
EDUCATION		
Graduate School	34.1	34.6
High School	40.3	40.2
Other	36.0	38.2
University	34.7	34.7

In-depth analysis (machine learning)

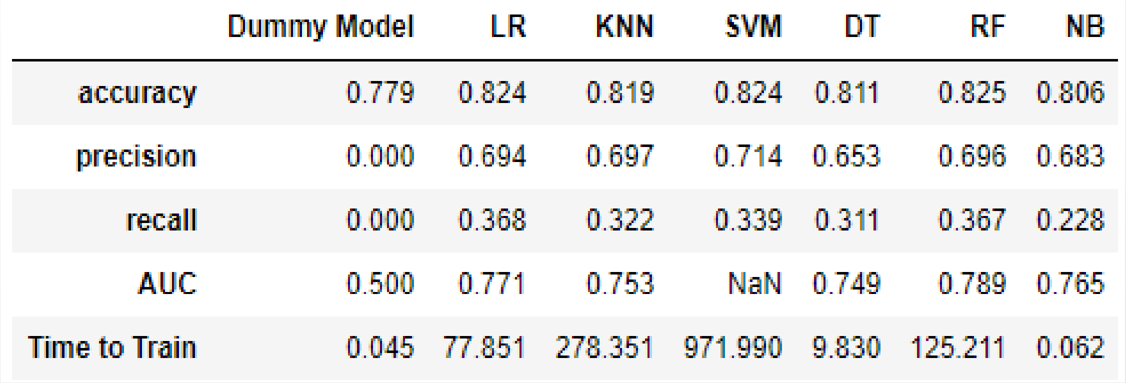
- 6 models were used for classification: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Naïve Bayes.
- Preprocessing the data:
  - Categorical features were preprocessed with sklearn's OneHotEncoder()
  - Numerical features were preprocessed with sklearn's MinMaxScaler()
  - Both wrapped in a ColumnTransformer() and put in a Pipeline() object
- Then we did train_test_split(), put our estimator (already in a pipeline) and hyperparameter space in a GridSearchCV() object, trained the model, and scored it.

In-depth analysis (machine learning)

- For scoring the models, we could not just use accuracy, because there was a 78-22 imbalance between no-default labels and default labels.
- Also, in our problem, false negative (predicting no default when default occurs) is much worse than false positives (the opposite).
- Therefore, a better metric than accuracy for us is recall: the ratio of correctly predicted defaulters to the number of actual defaulters.
- It is this that we aimed to maximize.

Analysis Results

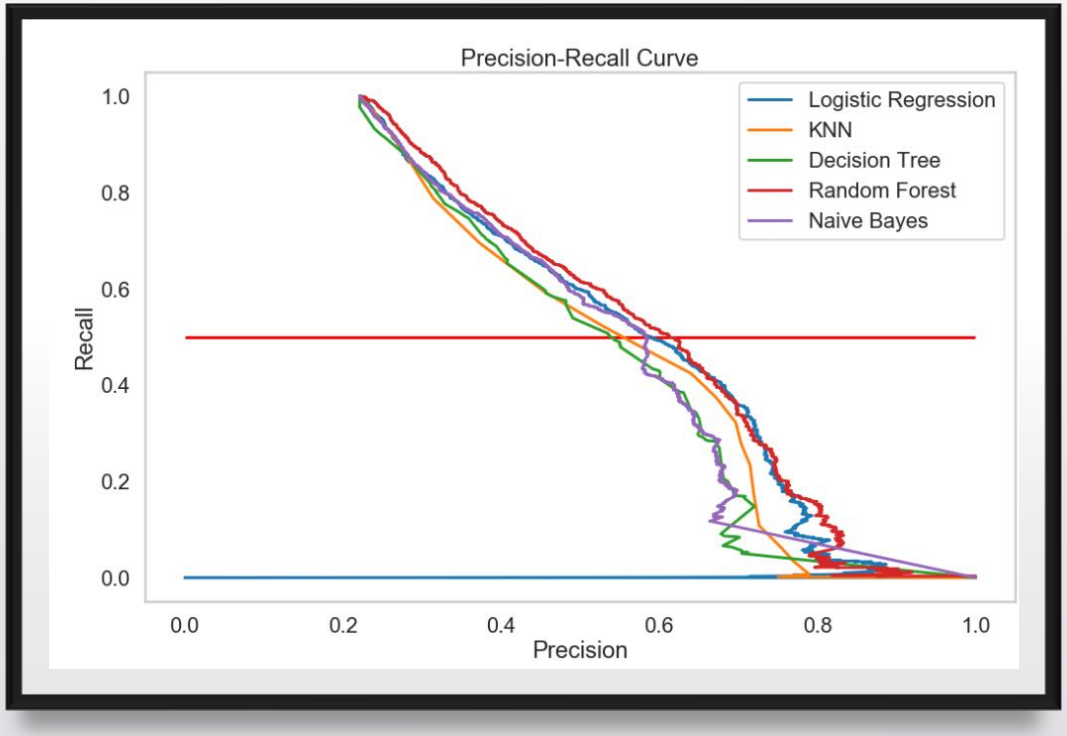
- All of our models performed better than the dumb model that always predicted no default.
- Also, by recall alone, Logistic Regression narrowly beat Random Forest.
- Not the whole story though, because the recall in the table is only the recall for one classification threshold: a threshold of 0.5.
- If we change the threshold, we will get different values of precision and recall.
- Can be seen in a precision-recall graph.



	Dummy Model	LR	KNN	SVM	DT	RF	NB
accuracy	0.779	0.824	0.819	0.824	0.811	0.825	0.806
precision	0.000	0.694	0.697	0.714	0.653	0.696	0.683
recall	0.000	0.368	0.322	0.339	0.311	0.367	0.228
AUC	0.500	0.771	0.753	NaN	0.749	0.789	0.765
Time to Train	0.045	77.851	278.351	971.990	9.830	125.211	0.062

Analysis Results (continued)

- When we look at the Precision-recall curves for each of the models, we see that Random Forest is overall the best (except for a small area where Logistic Regression is better)
- Therefore, if we want to maximize recall, we should use a Random Forest
- When we do that and use a threshold of 0.25, we get much higher recall (60.6%) without sacrificing too much accuracy (78.6%) or precision (51.5%)



Recommendations for Client

1. If you use one model to predict default, use a Random Forest because it achieves the highest recall score, which is the metric that matters most.
2. Because we saw in the Exploratory Analysis phase that the latest month’s data was most important for predicting default, greater effort should be made to ensure that this month’s data is accurate and if more data can be obtained about a client’s financial state in the month prior to possible default, predictions can be improved.

Suggestions for Improvement

1. Feature Engineering: consider interaction terms among the features.
2. More EDA: perhaps use clustering for better intuition of customer base.
3. Sampling techniques for class-imbalance: upsampling, downsampling, SMOTE.
4. More Advanced Models: Neural Networks, Discriminant Analysis, Gradient Tree Boosting, Ada Boost.
5. Unsupervised Learning: Dimensionality reduction for leaner models.