Credit Card Default Prediction

Spring 2023, Erdos Bootcamp

Team members: B. Mostowski, S. Provat, M. Ali, M. Molina, A. Lassoued

Overview

Problem - Banks when issuing credit cards undertake default risk, or the risk that consumers will fail to make payments on their debt. It is helpful for banks, hence, if possible, to detect ahead of time if a consumer will default on their payments.

Stakeholders - Banks, who wish to anticipate beforehand when a customer may default on their credit card payments and hence reduce risk and uncertainty with credit card issuance.

Task - Given the payment history of a consumer as well as personal characteristics (sex, education, marriage, age), predict whether the consumer will default on their payment or not.

Description of Data

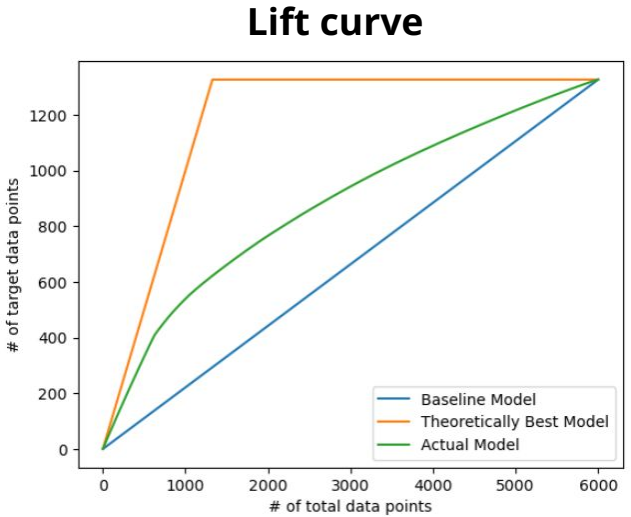
Our project collected payment data from a significant bank in Taiwan, specifically focusing on credit card holders. The data was donated in January 2016 and consisted of 30,000 observations. Out of these, 22.12% (6636 observations) represented cardholders who defaulted on their payments. To analyze the data, we used a binary variable, with “Yes” indicating default payment (1) and “No” indicating non-default payment (0). The data 23 explanatory variables:

- X1: Amount of the given credit (NT dollars)
- X2: Gender (Male = 1, Female = 2)
- X3: Education (graduate school = 1, university = 2, high school = 3, others = 4)
- X4: Marital status (married = 1, single = 2, others = 3)
- X5: Age in years
- X6 through X11: Past payments history from April to September 2005 with X6 = payment status in September 2005 through X11 = payment status in April 2005. (pay duly = -1, one month delay in payment = 1, going through nine or more months delay in payment = 9.)
- X12 through X17: Bill statement amount (NT dollars) with X12 = amount for September 2005 through X17 = amount for April 2005.
- X18 through X23: Previous payment amount (NT dollars) with X18 = payment for September 2005 through X23 = payment for April 2005.

Different Model Approaches

We considered several approaches to construct a prediction model for the data:

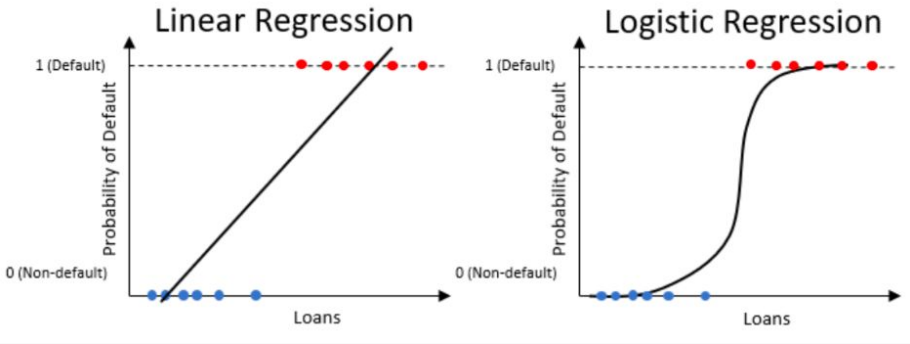
- Logistic regression
- Naive Bayes classifier
- Random forest classification
- K-nearest neighbors classification
- Neural networks



Comparison metric:

- Validation accuracy
- Lift curve:
  - Used to show how much better a predictive model is compared to a random guess
  - Lift curve area ratio=(area between the model curve and the baseline curve)/(area under the theoretically best model and the baseline curve)
  - Larger ratio means better predictive capabilities than random guess.

Logistic Regression



- Extended version of linear regression but only gives values between 0 and 1.
- Gives a straightforward probabilistic formula for classification.
- Uses the logistic function (called sigmoid function) to model the relationship between the independent variables and the probability of the event.

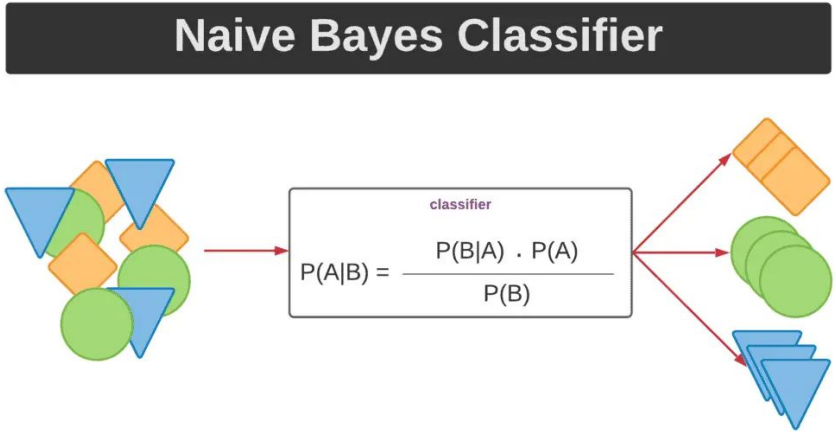
Advantages: simplicity, efficiency, and interpretability.

Limitations: assumes linearity and independence of observations

Naive Bayes

Naive Bayes is a popular classification algorithm based on Bayes' theorem.

- It assumes that features are conditionally independent given the class label.
- low computational complexity and can handle large high-dimensional datasets with a large number of features efficiently.
- widely used for spam filtering, sentiment analysis, and document categorization.

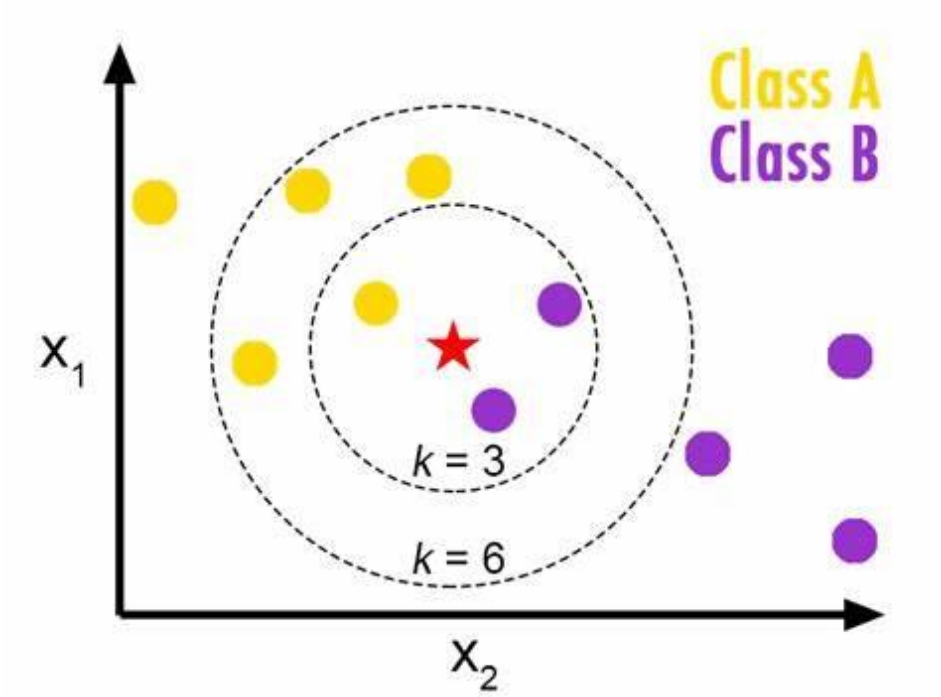


Limitations:

Feature independence is not always justified

K-Nearest Neighbors

- Main Idea: Points clustered together ought to belong to the same class
- **Advantages** - Intuitive, no assumptions, very easy to implement
- **Limitations** - Curse of dimensionality, also may not perform well for imbalanced data set.

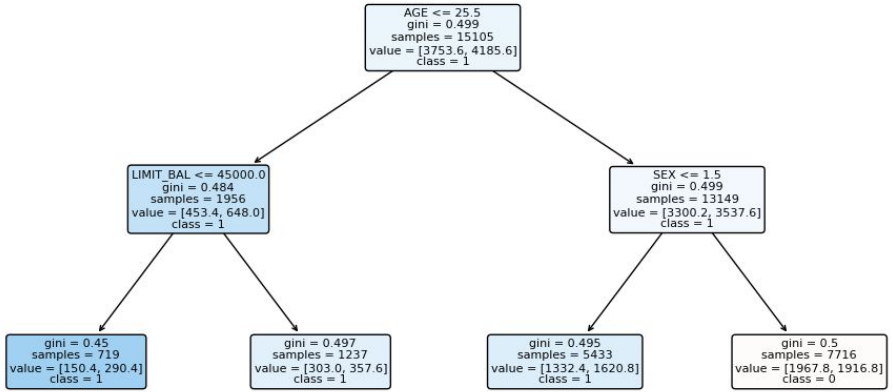


Random Forest Classification

This method is straightforward to interpret and flexible

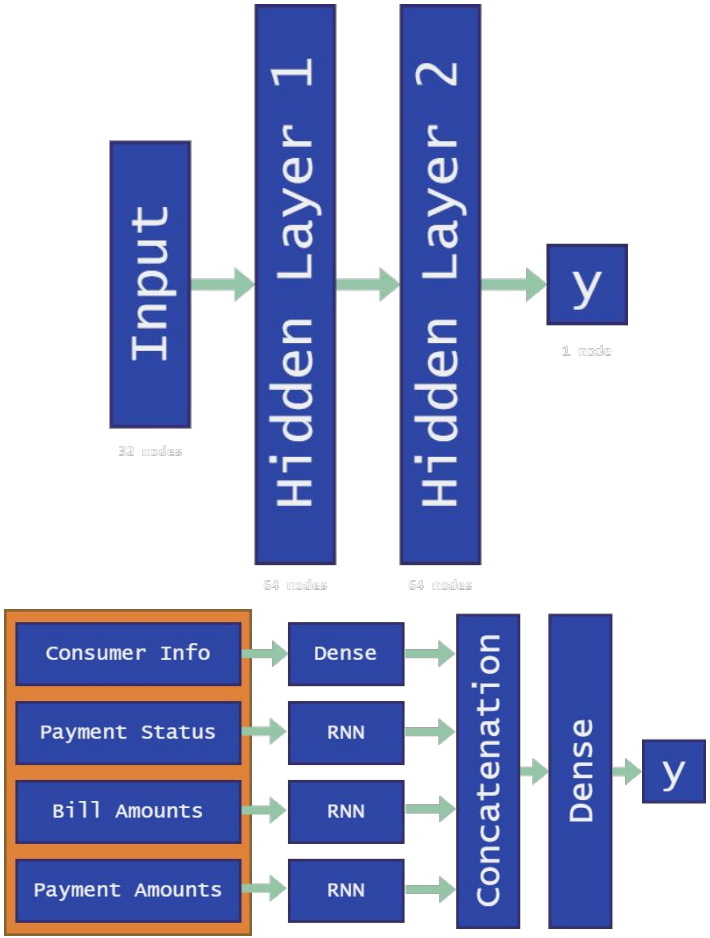
The classification of a customer is done through a sequence of decisions (tree) that depend on selected features (latest payment status, credit limit, age,...).

The class is determined using a voting system with a large number of randomly generated decision trees.



Neural Networks

- Neural networks work by pushing the input through multiple layers of nodes via linear algebra operations.
- We tried two different architectures on the problem (see right)
- **Advantage** - These networks are robust enough when trained to model nonlinear or complex relationships between input and output.
- **Disadvantage** - computationally expensive, black box nature, necessity for massive amounts of data



Project Takeaways

- To assess results across the different tested models, we examined both model accuracy and also a measure called “area ratio”, which gauges how relatively quickly the model makes accurate predictions.
- The Random Forest classifier only uses the payment statuses (inclusion of payment info led to overfitting)
- The Neural Network with partially recurrent structure performed *worse* than the simple neural network with dense layers, so results for the simple network are shown.
- K-nearest neighbors was conducted with K = 5

Model Type	Validation Accuracy	Area Ratio
Logistic Regression	81.5%	0.454
Random Forest	82.0%	0.416
K-nearest Neighbors	79.4%	0.741
Naive Bayes	63.6%	0.795
Neural Networks	80.3%	0.639

References

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.