Научная статья
Original article

Интеллектуальная оценка кредитных рисков при помощи алгоритмов машинного обучения в среде RStudio

Тыртышняя М.С.
Уральский государственный экономический университет, г. Екатеринбург, Россия
Автор-корреспондент: mari212000@mail.ru

Аннотация: В статье рассматривается применимость алгоритмов машинного обучения для оценки рисков кредитования физических лиц. Для оценки кредитных рисков используются алгоритмы машинного обучения Random Forest и XGBoost. Ключевым звеном в достижении поставленной цели выступает среда RStudio языка программирования R. Проводится анализ качества интеллектуальных моделей с помощью построения матрицы ошибок (Confusion Matrix) с последующим расчетом следующих метрик: точности (Accuracy). Результаты проведенного исследования могут быть использованы при анализе кредитных рисков при выдаче кредитов физическим лицам.

Ключевые слова: машинное обучение, интеллектуальная модель, анализ, кредитные риски, алгоритм.

Для цитирования: Тыртышняя М.С. Интеллектуальная оценка кредитных рисков при помощи алгоритмов машинного обучения в среде RStudio. Умная цифровая экономика. 2023. Т.3, №2, с. 24-29

Intelligent credit risk assessment using machine learning algorithms in RStudio environment

Tyrtyshnyaya M.S.
Ural State University of Economics, Yekaterinburg, Russia
Corresponding author: mari212000@mail.ru

Abstract: The article discusses the applicability of machine learning algorithms for assessing the risks of lending to individuals. Random Forest and XGBoost machine learning algorithms are used to assess credit risks. The key link in achieving this goal is the RStudio environment of the R programming language. The analysis of the quality of intellectual models is carried out using the construction of an error matrix (Confusion Matrix) with the subsequent calculation of the following metrics: accuracy (Accuracy). The results of the study can be used in the analysis of credit risks when issuing loans to individuals.

Keywords: machine learning, intellectual model, analysis, credit risks, algorithm.

For citation: Tyrtyshnyaya M.S Intelligent credit risk assessment using machine learning algorithms in RStudio environment. Smart Digital Economy. 2023. Vol. 3, №2, pp. 24-29

Потребительское кредитование – одна из наиболее прибыльных отраслей банковской деятельности. Однако потребительские кредиты связаны с повышенным риском невозврата денежных средств банкам, особенно это выражается на фоне роста кредитной нагрузки населения и усиления конкуренции между кредитными организациями. Проблема выявления банками рисков заемщиков в текущих условиях решается разработкой простых в реализации, но эффективно выявляющих потенциально рисковых заемщиков, моделей кредитного скоринга. В качестве альтернативы широко распространенным статистическим методам для построения скоринговых моделей могут быть применены алгоритмы машинного обучения [1].

Целью данной работы является исследование возможностей машинного обучения (при помощи алгоритмов машинного обучения Random Forest и XGBoost в среде RStudio языка программирования R [3]) для оценки кредитных рисков физических лиц.

Построение интеллектуальной модели для оценки кредитных рисков.

Для построения модели, необходимо определить предикторы. Предиктор (прогнозирующая переменная) – переменная, используемая для прогнозирования целевой переменной (Target Variable).

Исходный набор данных содержит 1000 записей с 10 качественными количественными предикторами. В этом наборе данных каждая запись представляет человека, который берет кредит в банке. Каждому человеку уже присвоен хороший или плохой кредитный риск (bad, good) (рисунок 1) [4].

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ Age             : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ Sex             : chr  "male" "female" "male" "male" ...
##  $ Job             : int  2 2 1 2 2 1 2 3 1 3 ...
##  $ Housing         : chr  "own" "own" "own" "free" ...
##  $ Saving.accounts : chr  NA "little" "little" "little" ...
##  $ Checking.account: chr  "little" "moderate" NA "little" ...
##  $ Credit.amount   : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ Duration        : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ Purpose         : chr  "radio/TV" "radio/TV" "education" "furniture/equipment" ...
##  $ Risk            : chr  "good" "bad" "good" "good" ...
```

Рисунок 1 – Предикторы модели

При разработке интеллектуальных моделей мы очистили данные и преобразовали их. Далее случайным образом разделили данные на обучающую и тестовую выборки (рисунок 2).

```
train<-sample(1:1000,800,replace = F)
gcrTrain<-gcr[train,]
```

Рисунок 2 – Разделение данных на выборки

Далее нами был проведен анализ предикторов, вычислили среднее значение и стандартное отклонение и построили график распределения плотности количественных переменных (рисунок 3).

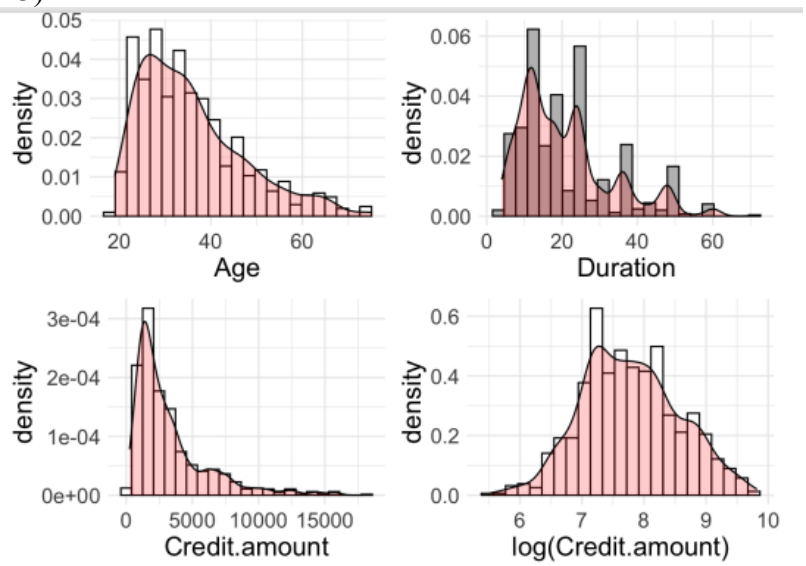

Рисунок 3 – Графики распределения плотности

Далее мы проанализировали качественные переменные. Вычислили относительную частоту (рисунок 4).

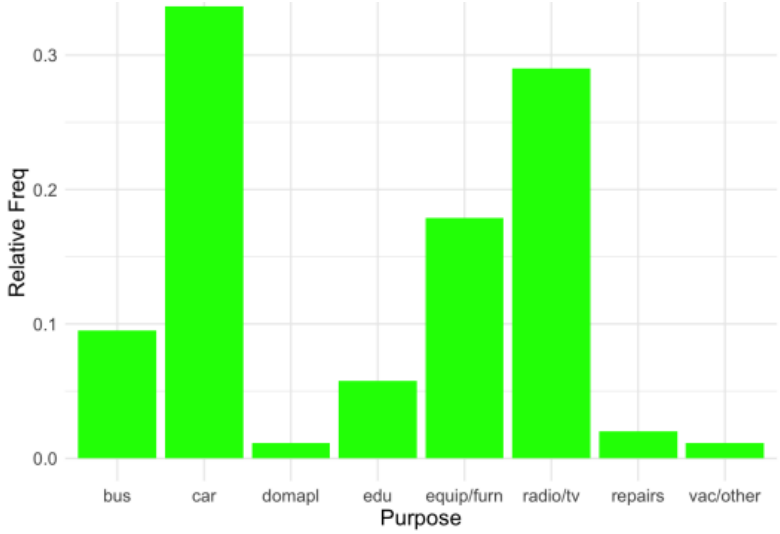

Рисунок 4 – Распределение целей кредита

На следующем шаге мы распределили переменны по группам риска (рисунок 5)

Sex	Risk		Job	Risk		Housing	Risk		Saving.accounts	Risk	
	bad	good		bad	good		bad	good		bad	good
female	0.33	0.67	UnSKNR	0.28	0.72	free	0.40	0.60	UnkOrNA	0.16	0.84
male	0.27	0.73	UnSKR	0.26	0.74	own	0.26	0.74	little	0.35	0.65
			SK	0.29	0.71	rent	0.34	0.66	moderate	0.30	0.70
			HSk	0.31	0.69				quite rich	0.14	0.86
									rich	0.15	0.85

Рисунок 5 – Распределение качественных переменных по группам риска

Из таблиц стоит отметить, что 73% мужчин имеют хороший кредитный риск, а у женщин данное значение равно – 67%. В следующей таблице 74% людей, имеющих собственный дом, имеют хороший кредитный риск. В последней таблице почти все люди с наибольшей суммой денег в банке имеют хороший кредитный риск.

Далее преобразуем данные, сберегательный счет, расчетный счет и место работы, так как они являются упорядоченными переменными. Преобразуем переменную house в фиктивные переменные, далее кодируем переменную purpose, используя 3-байтовую кодировку, например 001 = bus, 010 = car. Затем разделяем значения на 3 столбца, так что значения bus в целевой переменной состоят из 3 переменных со значениями 0, 0 и 1 соответственно (рисунок 6).

```
'data.frame':   800 obs. of  13 variables:
 $ Age             : num  -0.0628 -0.1501 -0.412 -0.4993 -1.2849 ...
 $ Sex             : num  0 0 1 1 1 1 0 0 1 1 ...
 $ Job             : num  2 3 3 2 2 3 2 2 2 2 ...
 $ Saving.accounts : num  0 1 1 1 2 1 1 0 1 1 ...
 $ Checking.account: num  1 2 0 2 2 2 0 1 0 0 ...
 $ Duration        : num  0.266 -0.714 -0.714 0.43 1.981 ...
 $ Risk            : num  0 1 1 1 0 1 1 1 1 0 ...
 $ logCredAmount   : num  -0.702 0.295 -0.254 1.534 0.301 ...
 $ Housing_own     : int  1 0 0 1 0 0 0 0 1 1 ...
 $ Housing_rent    : int  0 1 1 0 1 1 1 1 0 0 ...
 $ byte1           : num  0 1 1 0 1 1 1 1 1 1 ...
 $ byte2           : num  1 0 1 1 1 1 1 0 0 1 ...
 $ byte3           : num  0 1 0 0 0 0 0 1 1 0 ...
```

Рисунок 6 – Преобразованные данные

Следующим шагом мы проверили, есть ли корреляции между переменными, чтобы выполнить метод главных компонент (Principal component analysis).

Итак, нами были выбраны 2 алгоритма машинного обучения Random Forest и XGBoost, чтобы посмотреть, какая модель работает лучше всего (рисунок 7).

```
nr<-nrow(gcrTrainFinal)
briecvmat<-matrix(numeric(50),nrow = 10,ncol = 2)
gcrTrainFinal<-gcrTrainFinal[sample(1:nr,nr,F),]
for(k in 1:10){
 id<-(1+(k-1)*trunc(nr/10)):(k*trunc(nr/10))
 Nid<-setdiff(1:nr,id)

 rf<-randomForest(factor(Risk)~.,data=gcrTrainFinal[Nid,])
 predrf<-predict(rf,gcrTrainFinal[id,])
 briecvmat[k,1]<-sum(predrf==gcrTrainFinal[id,7])/length(predrf)



 xgb<-xgboost(as.matrix(gcrTrainFinal[Nid,-7]),gcrTrainFinal[Nid,7],objective="binary:logistic",nrounds = 150,verbose = 0)
 predxg<-ifelse(predict(xgb,as.matrix(gcrTrainFinal[id,-7]))>0.5,1,0)
 briecvmat[k,2]<-sum(predxg==gcrTrainFinal[id,7])/length(predxg)


}
apply(briecvmat, 2, mean)
```

```
## [1] 0.745 0.725
```

Рисунок 7 – Реализация моделей с алгоритмами машинного обучения

Стоит отметить, что обе модели показывают примерно равные результаты. Из этого следует, что необходимо оценить их качество при помощи построения матрицы ошибок (Confusion Matrix) с последующим расчетом точности (Accuracy) [1].

Первым шагом мы создали тестовую выборку, далее вычислили предсказанные значения и, наконец, построили для каждой модели матрицу ошибок по тестовой выборке с помощью функции confusionMatrix из пакета «caret» (рисунки 8 и 9).

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   0   1
##          0  28  15
##          1  42 115
##
##               Accuracy : 0.715
```

Рисунок 8 – Матрица ошибок для randomForest и Accuracy

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   0   1
##          0  34  20
##          1  36 110
##
##               Accuracy : 0.72
```

Рисунок 9 – Матрица ошибок для XGBoost и Accuracy

Как видно, точность (Accuracy) randomForest составляет 0,715, а точность модели с использованием XGBoost составляет 0,72.

Итак, модели были обучены на тренировочных данных, но не достигли идеальной производительности. Значения Accuracy, равное 0,715 и 0,720, указывают на то, что модели имеют некоторые ограничения в различении между классами и могут делать ошибки при классификации некоторых объектов. Модель с использованием XGBoost имеет более точные

результаты. Несмотря на то, что модели не идеальны, они все еще могут иметь практическую ценность для решения конкретных задач. Можно использовать ROC-кривую, чтобы определить оптимальный порог бинарной классификации для достижения наилучшей производительности модели на тестовых данных.

Таким образом, нами были разработаны модели машинного обучения для распознавания кредитных рисков, которые могут использоваться для принятия решений о выдаче кредитов.

Список литературы

1. Бруссард, М. Искусственный интеллект: пределы возможного / Мередит Бруссард; пер. с англ. - Москва: Альпина нон-фикшн, 2020. - 362 с. - Текст: электронный. - URL: https://znanium.com/catalog/product/1220958 (дата обращения: 20.04.2023).
2. Золотарюк, А. В. Язык и среда программирования R: учебное пособие / А.В. Золотарюк. — Москва: ИНФРА-М, 2023. — 162 с. — (Высшее образование). — Текст: электронный. - URL: https://znanium.com/catalog/product/2049696 (дата обращения: 23.04.2023).
3. Ланц Бретт Машинное обучение на R: экспертные техники для прогностического анализа. –СПб.: Питер, 2020. – 464 с.
4. Назаров Д.М. Data Science и интеллектуальный анализ данных : Учебное пособие / Д. М. Назаров, С. В. Бегичева, Д. Б. Ковтун, А. Д. Назаров. – Москва : Ай Пи Ар Медиа, 2023. – 304 с. – ISBN 978-5-4497-1931-7. – EDN VQKMUJ.
5. Назаров, Д. М. Формирование метапредметных компетенций в курсе "информационные технологии" средствами языка обработки больших данных R / Д. М. Назаров // Информатика и образование. – 2019. – № 4(303). – С. 12-22. – DOI 10.32517/0234-0453-2019-34-4-12-22. – EDN NBOAQY.