Моделирование дефолта на рынке автокредитования
Шарый А. А.

Шарый Алексей Александрович / Sharyj Aleksej Aleksandrovich – магистрант,
кафедра прикладной математики,
Федеральное государственное бюджетное образовательное учреждение высшего образования
Финансовый университет при Правительстве Российской Федерации, г. Москва

Аннотация: *статья посвящена оценке факторов, влияющих на дефолт заёмщика автокредита. В работе будет приведено сравнение нескольких методов моделирования вероятности дефолта: логистической регрессии, регрессии Кокса и дерева решений. Процесс моделирования будет осуществляться в RStudio – свободной среде разработки программного обеспечения с открытым исходным кодом для языка программирования R, предназначенной для статистической обработки данных и работы с графикой.*
Ключевые слова: *дефолт, автокредит, заёмщик, моделирование.*

Для проведения исследований получены данные одного из санированных банков.

По автокредитованию доступна, оказалась лишь следующая информация:

Пол заемщика (Gender): мужской, женский;

Возраст заёмщика (age): количество полных лет, от 27 до 66;

Город проживания (Adress): 1 – Москва или Санкт-Петербург, 0 – другой город;

Дата начала кредитования (BeginDate): 1 – выдача кредита в кризисный год (2008, 2009), 0 – в некризисный;

Сумма кредита (SumK): в тысячах рублей;

Срок кредита (Term):в месяцах, от 12 до 84;

Ставка по кредиту (Rate): от 9 % до 22 %;

Количество дней просрочки (Delay): согласно Базель II представлена двумя категориями – 1 (дефолт – более 90 дней просрочки) и 0 (нет дефолта).

Объем выборки составляет чуть более 7100 автокредитов. После исключения неполных данных итоговый объем выборки составил 1744 автокредита.

Логит-модель или логистическая регрессия – это статистическая модель, которая применяется для предсказания вероятности возникновения некоторого события по имеющимся значениям множества переменных. Причем переменные могут быть как количественными, так и категориальными. На практике логистическая регрессия обычно используется для решения двух задач: моделирования взаимосвязей и классификации данных. В нашем случае логистическая модель будет выглядеть следующим образом:

$$risk\ of\ default = \frac{1}{1+e^{-z}},$$

где $z = \beta_0 + \beta_1 Gender + \beta_2 age + \beta_3 Sum + \beta_4 rate.$

Регрессия Кокса относится к классу моделей выживания. Регрессия Кокса или, как её ещё называют, модель пропорциональных рисков – это модель, позволяющая прогнозировать риск наступления события с рассматриваемым объектом и оценивать влияние заранее определенных независимых переменных на риск наступления этого события. При этом риск рассматривается как функция, зависящая от времени. Так как риск – это не вероятность, то его значения могут превышать единицу.

В нашем случае модель регрессии Кокса выглядит следующим образом:

$$h_i(t) = h_0(t)\exp(\beta_1 Gender_{i1} + \beta_2 age_{i2} + \beta_3 Sum_{i3} + \beta_4 Term_{i4} + \beta_5 Rate_{i5}),$$

Для построения дерева решений сначала отыскивается переменная, наилучшим образом разбивающая корневой узел (весь набор данных) на два узла-потомка (подгруппы).

Процесс бинарного расщепления повторяется для каждого узла до тех пор, пока возможно улучшение прогноза.

Для отбора предикторов и разбиения узлов метод CART применяет принцип уменьшения остаточной суммы квадратов, называемой девиацией или уклонением. На каждом этапе выбирается расщепление, дающее максимальное уменьшение девиации.

До начала расщепления девиация имеет вид:

$$D = \sum_j (y_j - \bar{y})^2.$$

В результате расщепления получается два подмножества – j_1 и j_2.

Девиация для любого расщепления равна:

$$D = \sum_j (y_j - \bar{y})^2 = \sum_{j_1} (y_{j_1} - \overline{y_1})^2 + \sum_{j_2} (y_{j_2} - \overline{y_2})^2 + n_1(\overline{y_1} - \bar{y})^2 +$$

$$+n_2(\overline{y_2} - \bar{y})^2.$$

Расщепление производится так, чтобы внутригрупповая сумма квадратов была как можно меньше, а межгрупповая как можно больше.

Загружаем данные в RStudio:

```
read.csv2("D:/Data/Autodata_wCity.csv")
data<-read.csv2("D:/Data/Autodata_wCity.csv")
```

Укажем тип переменных:

```
data$Gender <-as.factor(data$Gender)
data$age<-as.numeric(data$age)
data$Adress<-as.factor(data$Adress)
data$Status<-as.factor(data$Status)
data$BeginDate<-as.factor(data$BeginDate)
data$SumK<-as.numeric(data$SumK)
data$Term<-as.numeric(data$Term)
data$Rate<-as.numeric(data$Rate)
data$Delay<-as.factor(data$Delay)
```

Здесь numeric – количественные переменные, factor – категориальные.

Построим модели с помощью следующих команд:

```
library(rpart)
library(pROC)
library(survival)
model<-rpart(Delay~., method='class', data, cp=0.001)
model2 <- glm(Delay ~Gender+age+SumK+Rate,family=binomial(link='logit'),data=data)
data(package = "survival")
read.csv2("D:/Data/Autodata_Surv.csv")
data2<-read.csv2("D:/Data/Autodata_Surv.csv")
data2$SurvObj <- with(data2, Surv(Time, Delay == 1))
res.cox1 <- coxph(SurvObj ~ Gender+age+SumK+Term+Rate, data =  data2)
```

За вывод полных результатов по любой модели отвечает команда summary.

Для оценки качества бинарной классификации часто применяют ROC-анализ. Построим ROC-кривые (кривые ошибок) для каждой модели.

```
roc.model<-roc(data$Delay,predict(model, type="prob")[1:1744],ci=TRUE)
plot.roc(roc.model)
roc.model2<-roc(data$Delay,predict(model2, type="response"),ci=TRUE)
plot.roc(roc.model2)
CoxRoc<-roc(data2$Delay,predict(res.cox1, type="risk")[1:1744],ci=TRUE)
plot(CoxRoc)
```

В результате будут выведены параметры каждой кривой ошибок. Определяющим параметром является площадь под графиком – AUC (area under the curve). Значение AUC близкое к 0,5 соответствует случайному гаданию и демонстрирует непригодность метода. Сведу основные значения в таблице 1:

Таблица 1. Показатели эффективности моделей

Модель/показатель	AUC	95 % доверительный интервал
Регрессия Кокса	0,5233	0,496-0,5507
Логистическая регрессия	0,6007	0,5741-0,6272
Дерево решений	0,8598	0,842-0,8776

К моему удивлению, модель, от которой я ожидал самых точных результатов – регрессия Кокса, оказалась самой неэффективной.

Тем не менее, довольно простая модель дерева решений, показала замечательные результаты. Соотношение между долей объектов от общего количества, верно классифицированных, как несущих признак и долей объектов, не несущих признака, но классифицированных как несущих признак здесь равно 0,8598.

Литература

1. *Груздев, А. В.* Метод бинарной логистической регрессии в банковском скоринге / А. В. Груздев // Риск-менеджмент в кредитной организации. — 2013. — № 2 (10). — С. 22-38.
2. *Agarwal S., Ambrose B. W., Chomsisengphet S.* Determinants of automobile loan default and prepayment. Economic Perspectives, 3Q, 2008, pp. 17-29.

3. *Agarwal, S., Ambrose B. W., Chomsisengphet S.* Information asymmetry and the automobile loan market in Household Credit Usage: Personal Debt and Mortgages. New York: Palgrave Macmillan, 2007, pp. 93–116.
4. *Calhoun, C. A., Deng Y.* A dynamic analysis of fixed- and adjustable-rate mortgage terminations. Journal of Real Estate Finance and Economics, Vol. 24, Nos. 1–2, January, 2002, pp. 9–33
5. *Clover M.* Yeh, Tsun-Siou Lee, The role of credit card behavior in auto loan grant decision. An application of survival table. Banks and Bank Systems, Volume 8, Issue 1, 2013. Pp. 112-117.