ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТОВ SAS ДЛЯ ОЦЕНКИ РИСКОВ ЗАЁМЩИКОВ

А.С. Вершинин
Научный руководитель: Е.И. Губин
Томский политехнический университет
vershinintmsk@gmail.com, gubine@tpu.ru

Введение

В связи с возрастающей необходимостью в автоматизации и повышении качества оценки кредитоспособности заемщика и его дальнейшего поведения, современные скоринговые модели усовершенствуются благодаря использованию методов интеллектуального анализа данных и машинного обучения. Сегодня эти методы уже не являются чем-то новым, это обязательный пункт банковских бизнес-процессов.

Скоринговая модель – это статистическая модель для прогноза вероятности попадания клиента в категорию «хороший» / «плохой» в течение периода времени после выдачи кредита, в течение которого определяется поведение заемщика по данному кредиту.

К интеллектуальному анализу данных принадлежит большое количество аналитических методов, которые обычно делятся на две большие категории: поиск закономерностей и прогнозное моделирование. SAS Enterprise Miner содержит множество методов и инструментов для поиска закономерностей и прогнозного моделирования. Программный продукт SAS Enterprise Miner (разработчик SAS Institute Inc.) - это интегрированный компонент системы SAS, созданный специально для выявления в огромных массивах данных информации, которая необходима для принятия решений. Разработанный для поиска и анализа глубоко скрытых закономерностей в данных SAS, Enterprise Miner включает в себя методы статистического анализа, соответствующую методологию выполнения проектов Data Mining (SEMMA) и графический интерфейс пользователя.

В данной работе для построения скоринговой модели рассматриваются такие статистические модели, как логистическая регрессия и деревья решений. Для построения модели были отобраны данные о заемщиках на основе анкетных данных.

Построение скоринговой модели подразумевает простой тип прогнозирования – решение или классификация. Такое прогнозирование обычно связано с категориальной переменной, что соответствует задачи принятия решения о заемщике.

Для подготовка исходных данных были использованы инструменты замены и импутации данных. Все интервальные переменные, значения которых отличались от среднего значения этой переменной более чем на три стандартных отклонения, были заменены на пропущенные значения. Кроме этого, были объединены разные уровни некоторых категориальных входных переменных. Далее интервальные входные переменные содержащие пропущенные значения, были заменены на среднее по всем непропущенным значениям этой переменной.

Такой подход исключает проблему неполных наблюдений. Любые изменения обучающий данных также распространяются на проверочные данные и другие данные из той же генеральной совокупности. Модель, построенная на измененных обучающих данных, не является смещенной, если те же изменения сделаны для любого другого набора данных, который обрабатывается этой моделью.

В ходе работы было построено дерево решений, где наблюдения оцениваются с помощью правил прогноза. Алгоритм поиска разбиения упрощает выбор входных переменных, а сложность модели управляется «обрубкой» дерева решений. На следующем рисунке изображено построенное дерево решений.

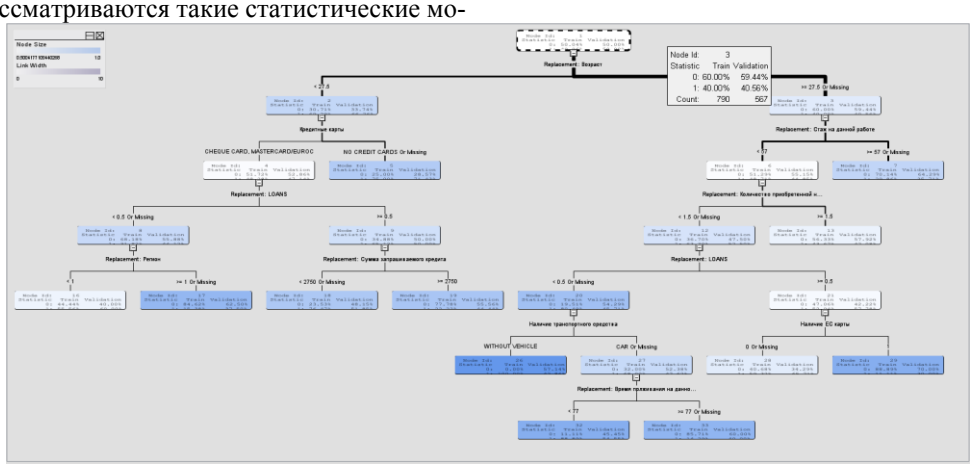

Рис. 1. Дерево решений

Основываясь на степени важности входных переменных и вероятностях, определенных алгоритмом дерева решений уже можно делать некоторые выводы по процессу принятия решения о заемщиках («хороший»/ «плохой»).

Для построения скоринговой карты была использована модель логистической регрессии с использованием метода пошагового добавления-удаления переменных. Результат построения скоринговой карты можно увидеть на рисунке 2.

Scorecard

		Scorecard Points
Imputed: Replacement: Возраст (IMP_REP_AGE)	IMP_REP_AGE< 25	-1
	25<= IMP_REP_AGE< 28	5
	28<= IMP_REP_AGE< 33	18
	33<= IMP_REP_AGE< 45, _MISSING_	25
	45<= IMP_REP_AGE	34
Imputed: Replacement: Количество приобретенной недвижимости (IMP_REP_PERS_H)	IMP_REP_PERS_H< 2, _MISSING_	10
	2<= IMP_REP_PERS_H< 3	22
	3<= IMP_REP_PERS_H< 4	23
	4<= IMP_REP_PERS_H< 5	20
	5<= IMP_REP_PERS_H	19
Imputed: Replacement: Стаж на данной работе (IMP_REP_TMJOB1)	IMP_REP_TMJOB1< 18	7
	18<= IMP_REP_TMJOB1< 60, _MISSING_	14
	60<= IMP_REP_TMJOB1< 144	23
	144<= IMP_REP_TMJOB1< 240	29
	240<= IMP_REP_TMJOB1	44
Replacement: Кредитные карты (REP_CARDS)	NO CREDIT CARDS, OTHER CREDIT CAR, _MISSING_, _UNKNOWN_	14
	CHEQUE CARD, MASTERCARD/EUROC	25
Replacement: Общий доход (REP_INCOMETOTAL)	REP_INCOMETOTAL< 3500	25

Рис. 2. Скоринговая карта

Заключение

В результате построения модели была произведена оценка эффективности модели по точности или ошибке классификации, прибыли или убыткам, и по статистике Колмагорова-Смирнова (KS). Точность и ошибка классификации подсчитывают правильные и неправильные прогнозы типа решений. В результате были получены следующие оценочные статистики: ошибка классификации – 0,36; KS – 0,3; коэффициент Gini – 0,37; Roc индекс – 0,69.

Последующими задачами данной работы являются оптимизация сложности регрессии, построение новых моделей, сравнение моделей и их применение.

Список использованных источников

1. Построение скоринговых карт с использованием модели логистической регрессии [Текст] // Интернет журнал «Науковедение» Выпуск 2, март - апрель 2014.

2. Особенности применения методов Data Mining в скоринговых решениях для коммерческих банков [Текст]// Журнал «Научные записки молодых исследователей» №3 – 2017 – С.5.

3. Прикладная аналитика с SAS Enterprise Miner [Текст]/ SAS Institute Inc. – 2015.