南京大學

金 融 系 统 仿 真

学 期 报 告

院　　系 工程管理学院

专　　业 工业工程

题　　目 Credit Scoring using Random Forests

年　　级 2010 级　学　　号 101279025

学生姓名 刘威志

指导老师 瞿慧　职　　称 副教授

报告提交日期 2013 年 12 月 27 日

南京大学金融系统仿真学期报告

学期报告题目：Credit Scoring using Random Forests

工程管理学院 院系 工业工程 专业 2010 级本科生 姓名：刘威志

指导教师：瞿慧 副教授

摘 要

贷款审批是普通商业银行的一项重大职能环节，在银行风险管理中占据着极其重要的角色。银行通过对借款者的信用进行评估，便可以更加准确的判断是否批准贷款，以及以多高的利率、多大的款额借给需求方，实现收益管理。因此，对于银行来说，判断借款者未来是否会违约以及违约大小便成了一个非常重要的问题。

本文通过互联网上两组贷款者申请贷款的历史信息记录及其最终违约与否的数据集，结合逻辑回归、分类树和随机森林这三种方法对借款者未来违约的概率进行预测，并通过准确率，KS 统计量，AUC 等对三种模型的有效性进行了对比。结果显示随机森林有效性最高，其次是逻辑回归，最差是分类树。

关键词： 信用评估；信用风险；逻辑回归；分类树；随机森林

目　　录

第一章　基本介绍

1.1　数据集

本文通过互联网收集了两组关于借款人申请贷款的相关信息数据集，其基本介绍见表1.1。

表 1.1　数据集介绍

数据集来源	数据集名称	数据集大小	违约比例
UCI machine learning repository	German Credit	1000	30%
Kaggle	Give me Some Credit	250,000	6.7%

1.1.1　German Credit

数据集 German Credit 的每一条记录包含 20 个特征变量（代表借款人为了从银行借款所提供的个人信息）以及 1 个分类变量（表示最终借款人是否发生了违约）。具体 20 个特征变量的介绍见表1.2

1.1.2　Give Me Some Credit

数据集 Give me Some Credit 的每一条记录包含 10 个特征变量（代表借款人为了从银行借款所提供的个人信息）以及 1 个分类变量（表示最终借款人是否发生了违约）。具体 10 个特征变量的介绍见表1.3

表 1.2　German Credit 数据集特征变量介绍

特征变量	解释
checking	status of existing checking account
duration	duration in month
history	credit history
purpose	purpose (e.g. car, furniture, repairs, etc.)
amount	credit amout
savings	savings account/bound
employ	present employment since
installment	installment rate in percentage of disposable income
status	personal status and sex
others	debtors/guarantors
residence	present residence since
property	property (e.g. real estate, life insurance, car etc.)
age	age in years
otherplans	other installment plans
housing	housing (e.g. rent, own, for free)
cards	number of existing credits at this bank
job	job types
liable	number of people being liable to provide maintenance for
tele	telephone
foreign	foreign worker

表 1.3　Give me Some Credit 数据集特征变量介绍

特征变量	解释
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits
Age	Age of borrower in years
NumberOfTime30-69DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years
DebtRatio	Monthly debt payments, alimony, living costs divided by monthly gross income
MonthlyIncome	Monthly income
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage and) and Lines of credit (e.g. credit cards)
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)

第二章　机器学习方法

首先，我们对原始数据集随机取样其 60% 记录作为我们的训练集，其余 40% 作为测试集。然后，分别利用逻辑回归、分类树以及随机森林对于训练集进行监督式学习，得到相应的分类器。

2.1　逻辑回归

2.1.1　逻辑回归简介

传统的线性回归 $y = \theta x$（其中 θ 为特征变量的系数行向量，x 为特征变量列向量）所得到的因变量 y 的取值范围往往在整个实数空间内，而一般的分类问题只有可数个因变量取值。逻辑回归利用逻辑函数 $h(z) = \frac{1}{1+e^{-z}}$（其中 $z = \theta x$）将实数空间内的数映射到 [0, 1] 上，使得连续问题转化为离散问题的 “概率” 解（见图2.1）。

2.1.2　逻辑回归算法

其合理性是由于在分界线（面，见图2.2）上的数据点的 z 值为 0，通过逻辑函数映射到 0.5，代表分解线（面）上的数据点无法判断其类别，可以理解为有 50% 的概率为 Negative Class，有 50% 的概率为 Positive Class；分界线（面）

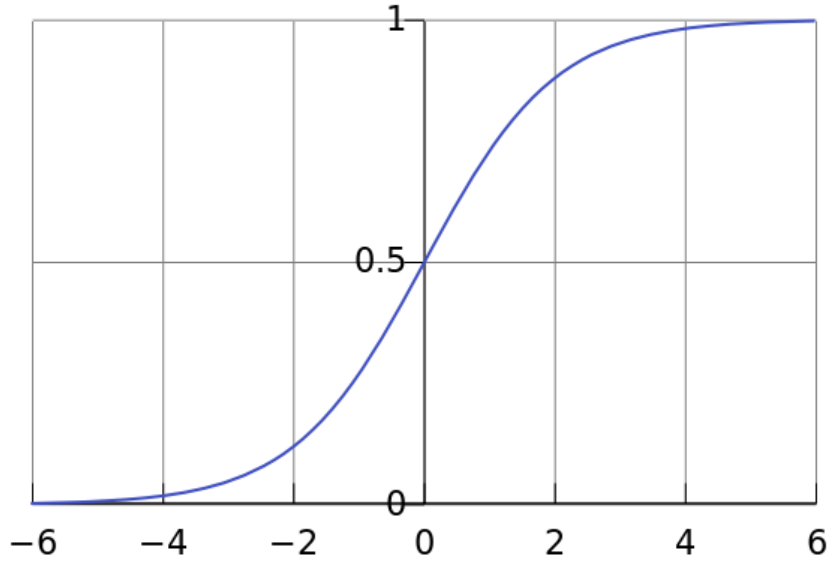

图 2.1　逻辑函数

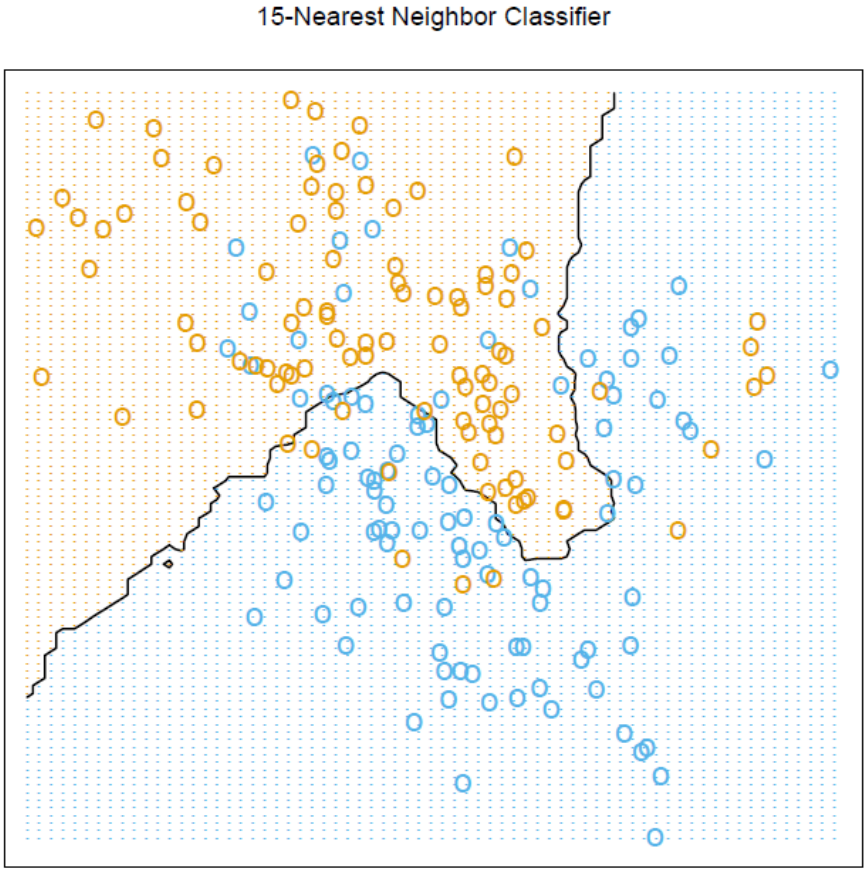

图 2.2　分界线

把多维空间分为了两个部分，其中正向空间内离分界线（面）越远 z 值越大，通过逻辑函数映射到 1，代表正向空间内远离分界线（面）的数据点有很大的可能性为 Positive Class。

具体的特征向量系数 θ 通过最小化成本函数 $cost(\theta) = ylog(h(\theta x)) + (1 - y)log(1 - h(\theta x))$ 便可获得，这样就得到了逻辑回归分类器。

2.1.3　逻辑回归结果

这里以 German Credit 的数据为例，表2.1显示了逻辑回归特征变量系数及其显著性。

除此之外，对于拒绝给予贷款的申请人我们给出了三项最有可能导致被拒的原因，见图2.3

表 2.1　逻辑回归结果

	Estimate	Std. Error	z value	$Pr(\geq \|z\|)$
(Intercept)	-5.48743	1.387014	-3.95629	7.61E-05 ***
checking	0.484832	0.089609	5.41055	6.28E-08 ***
duration	-0.01426	0.01144	-1.24679	0.212473
history	0.469549	0.114837	4.088823	4.34E-05 ***
purpose	0.068907	0.042475	1.622316	0.104736
amount	-0.00015	5.07E-05	-2.99869	0.002711 ***
savings	0.25071	0.076253	3.287871	0.001009 ***
employed	0.109647	0.095728	1.145409	0.25204
installp	-0.33189	0.106943	-3.10348	0.001913 ***
marital	0.48904	0.155785	3.139197	0.001694 ***
coapp	0.100784	0.225263	0.447406	0.654582
resident	-0.10415	0.101927	-1.02184	0.306856
property	-0.14986	0.12051	-1.24353	0.213674
age	0.019083	0.011009	1.73332	0.083039 *
other	0.325128	0.140454	2.314837	0.020622 **
housing	0.274481	0.216319	1.268869	0.204488
existcr	-0.39916	0.203087	-1.96544	0.049364 **
job	0.34085	0.179191	1.902159	0.057150 *
depends	-0.24825	0.290759	-0.8538	0.393214
telephon	0.021655	0.244287	0.088645	0.929364
foreign	1.699978	0.891746	1.906348	0.056605 *

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	age	other	housing	existcr	job	depends	telephon	foreign	good_bad	score1	score2	score3	topk	
2	67	3	2	2	3	1	2	1	good	0.947141	0.785714	0.846	savings;history;age	
3	22	3	2	1	3	1	1	1	bad	0.446593	0	0.326	installp;property;existcr	
4	53	3	2	1	3	1	1	1	good	0.931717	0.963235	0.982	checking;age;savings	
5	61	3	2	1	2	1	1	1	good	0.906865	0.963235	0.926	checking;age;savings	
6	28	3	2	2	4	1	1	1	bad	0.67592	0.666667	0.674	history;marital;job	
7	24	3	1	1	3	1	1	1	bad	0.242119	0	0.272	purpose;existcr;other	
8	28	3	1	1	3	1	1	1	good	0.398376	0.3	0.266	installp;amount;existcr	
9	25	1	2	3	3	1	1	1	good	0.145249	0.266667	0.466	savings;purpose;installp	
10	44	3	3	1	4	1	2	1	bad	0.298514	0.785714	0.276	job;housing;employed	
11	44	3	1	1	3	2	1	1	good	0.765885	0.7	0.814	installp;savings;duration	
12	44	3	2	1	3	1	1	1	good	0.889959	0.7	0.868	history;amount;age	
13	39	3	2	1	2	1	1	1	good	0.779685	0.963235	0.766	checking;marital;amount	
14	63	3	2	2	3	1	2	1	bad	0.483161	0.2	0.378	age;purpose;history	
15	30	3	2	2	3	1	2	1	good	0.558439	0.3	0.612	installp;marital;other	
16	25	3	2	2	2	1	1	1	bad	0.45848	0	0.398	history;marital;other	
17	30	1	2	1	4	1	1	1	good	0.839057	0.782609	0.858	checking;job;savings	
18	26	3	2	1	3	1	2	1	good	0.647998	0.7	0.672	amount;savings;resident	

图 2.3　导致贷款被拒最有可能的三项原因

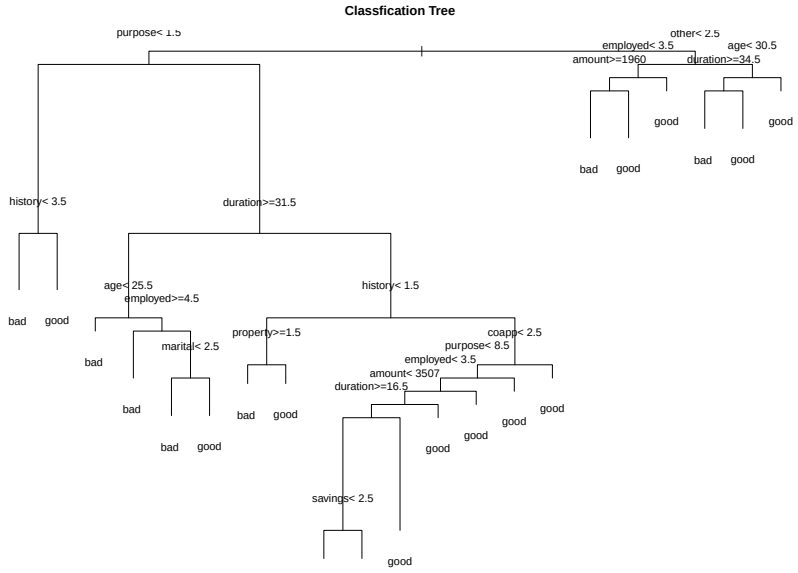

图 2.4　分类树结果

2.2　分类树

2.2.1　分类树简介

分类树通过对数据集按照不同的标准进行分枝，直至无法再分为止。其叶结点表示最终分类结果，其余节点表示各个分类逻辑表达式。分类树建立的难点在于如何选取分类标准以及相应阈值。传统的 ID3 是使用信息熵来刻画数据集的纯度，并选择使得信息增益最大的分类标准作为非叶结点的分裂标准。

2.2.2　分类树算法

数据集 S 的纯度可以用其信息熵来刻画：$H(S) = -\sum_i p(i)log(p(i))$，其中 $p(i)$ 表示种类 i 出现的频率，可见若数据集 S 中的数据只有一类的话，信息熵达到最小值 0。其次，非叶节点的分裂标准通过选择使得信息增益 $IG(Y) = H(S) - \sum_{i\in T} p(i)log(p(i))$ 最大的分类标准（式中子数据集 T 表示按此分类标准得到的子数据集）。

2.2.3　分类树结果

分类树的结果如图2.4，因为图像显示问题，这里的根节点应为 $checking \leq 2.5$。

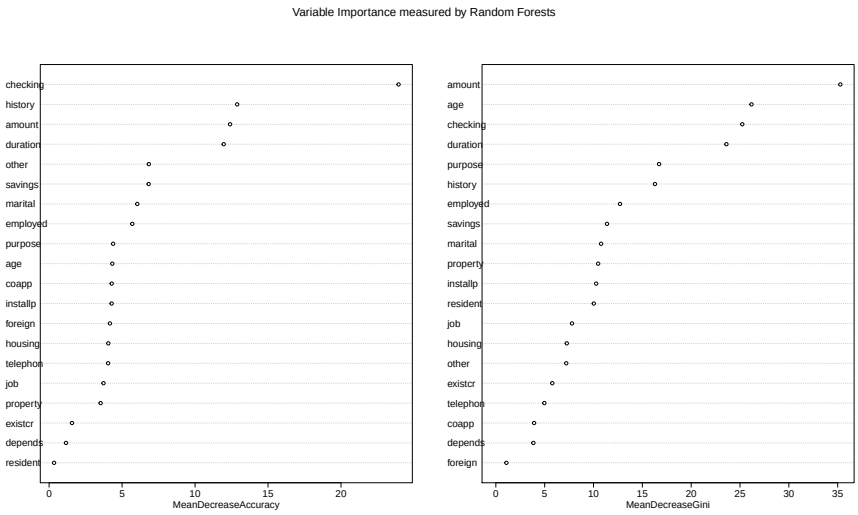

图 2.5　特征变量重要性

2.3　随机森林

2.3.1　随机森林简介

随机森林是一种集成式学习方法，包含了多个分类树分类器，最终的分类结果由众多分类器投票得出。随机森林因为其计算效率高、适用于并行处理、可以处理大量数据、评估变量重要性，在最近几年机器学习领域比较流行。微软的体感游戏设备 Kinect 也利用了随机森林来识别玩家的动作。

2.3.2　随机森林算法

随机森林是以分类树为基础的集成式学习方法，与其相比，主要有两大区别：

- 每棵分类树所使用的数据集是通过自助法对原始数据集进行有放回重取样得到的。
- 每棵分类树训练所用的特征变量是随机给定的。

最终的分类结果由各棵分类树投票得出。

2.3.3　随机森林结果

随机森林得到的特征变量重要性如图2.5

第三章　模型评估

3.1　模型评估标准

我们选取了准确率，KS 统计量以及 AUC 作为三种模型的评估标准，为了理解这三种评估标准，首先介绍一下分类问题的最终可能几种结果，见表**??**。

	真实值	
预测值	真阳 (TP)	伪阳 (FP)
	伪阴 (FN)	真阴 (TN)

其中准确率等于 $\frac{TP+TN}{TP+FP+FN+TN}$，KS 统计量以及 AUC 是通过接受者操作特征曲线（ROC 曲线）得到的。ROC 曲线的横坐标为 $FalsePositiveRate = \frac{FP}{FP+TN}$，纵坐标为 $TRUEPositiveRate = \frac{TP}{TP+FN}$（见如3.1）。其中 KS 统计量为纵坐标与横坐标差值的最大值，AUC 为 ROC 曲线与横坐标所围的面积，其理想值为 1。

3.2　模型评估结果

3.2.1　German Credit

以 German Credit 为训练数据，得到以下结果：

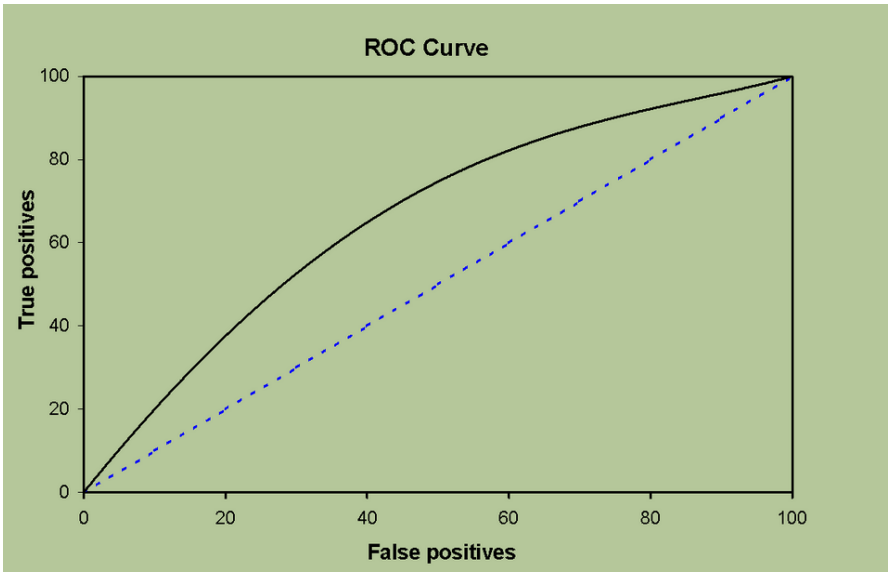

图 3.1　接受者操作特征曲线（ROC）

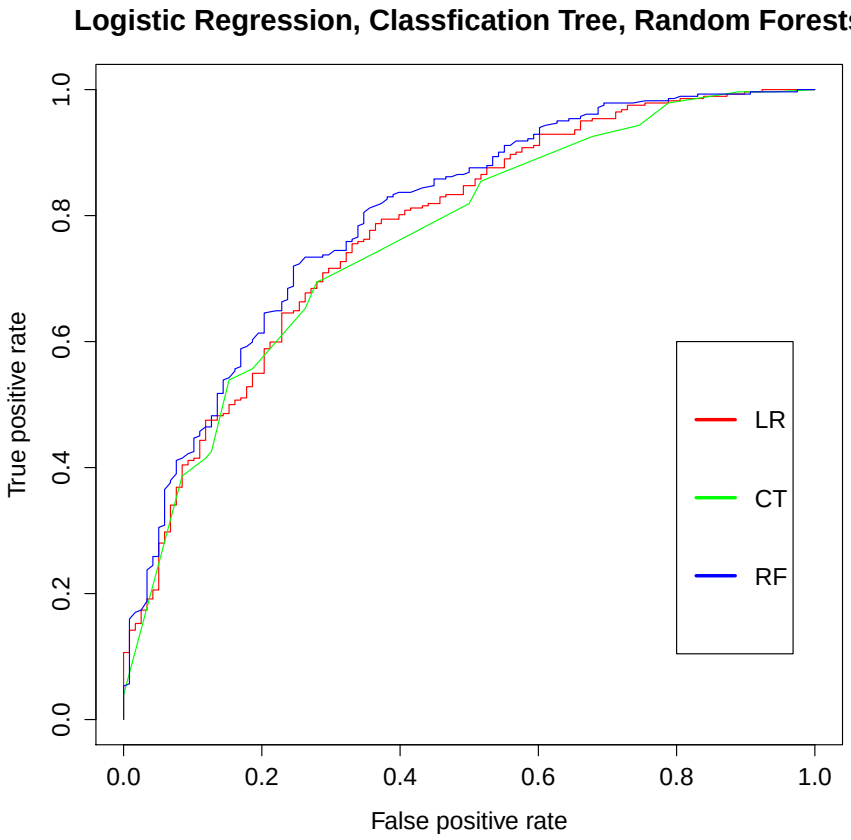

图 3.2　German Credit - ROC

表 3.1　German Credit

Model	KS	AUC	Accuracy	Cutoff
LR	0.425	0.776	0.773	0.421
CT	0.415	0.762	0.753	0.250
RF	0.474	0.798	0.780	0.472

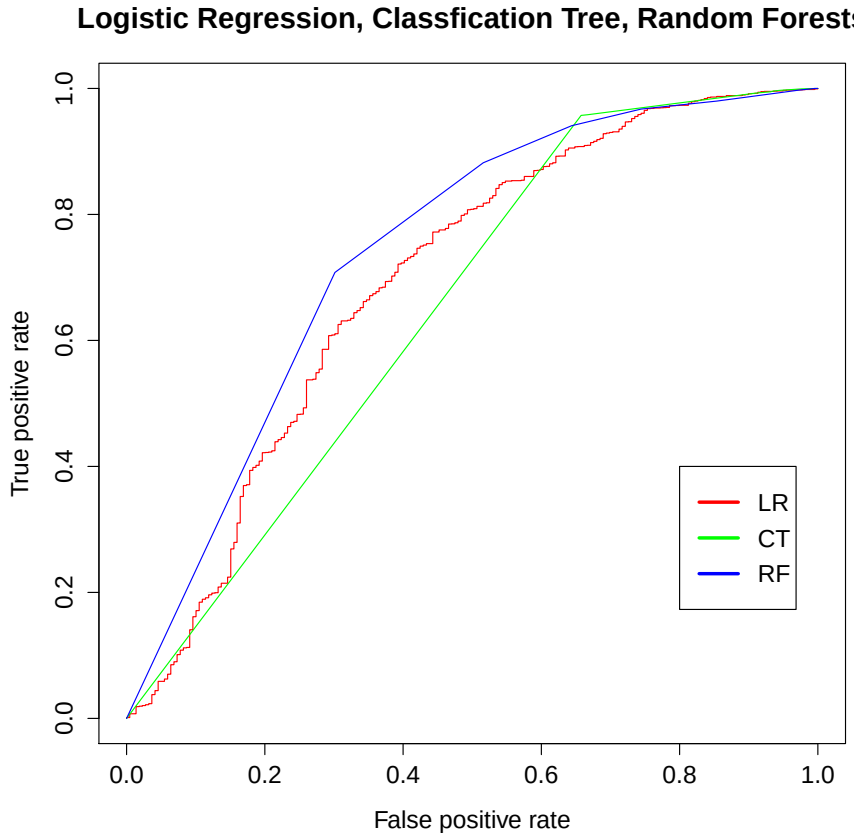

图 3.3　Give Me Some Credit - ROC

3.2.2　Give Me Some Credit

以 Give Me Some Credit 为训练数据，得到以下结果：

表 3.2　Give Me Some Credit

Model	KS	AUC	Accuracy	Cutoff
LR	0.329	0.695	0.933	0.727
CT	0.300	0.651	0.933	0.308
RF	0.407	0.741	0.932	0.300

第四章　结论

通过以上模型评估结果我们可得到如下结果：

- 随机森林的效果最佳，其次是逻辑回归，分类树表现最差。
- 随机森林与分类树的对比可以看出集成式学习的优势。
- 随机森林可以通过和逻辑回归结合起来以提高其效果。