Titanicの乗客データを用いた分析

３年　川瀬　宮川　渡邉

演習目的と内容

データ分析において、より適合性の高い分析モデルを探索し、作ることができるようになること。

今回はTitanic号海難事故(1912)の生存者に関するデータを用いて、生存者が生き残った要因を探るべく、まず分析モデルを複数構築した。その後、予測用データにそのモデルを適用して実際の生存者のデータと照らし合わせ、その正答率のより高いモデルを採用することでモデルの適合性をより高め、予測力の高いモデルを探った。

背景知識

事故発生当時の20世紀初頭、「世界一安全で決して沈まない豪華客船」と謳われたTitanic号は、2223人の乗客を乗せ、4月10日にサウザンプトンからNYに向けて出航した。

しかし、14日の（原因は諸説はあるが）無風で海面に波が立たず、進路前方に迫る氷山の発見が遅れたため、Titanic号は氷山を回避しきれず接触、15日未明にニューファンドランド島沖で沈没した。

タイタニック号には乗客全員が避難出来るだけの数のボートは搭載されていなかったため、1500人を超える犠牲者を出した。

避難の際、船員は“Women and children first”、“Be British, boys, be British.”と女性や子供を優先的に避難させるように誘導したと言われており、その英国紳士的な振る舞いは評価されている。

また、女性や子供の避難が終わった後は、グレードの高い一等客室の客から避難ボートに誘導されたようである。

今回の演習では、これらの史実を踏まえながら分析モデルを構築し、検証する。

分析方法

被説明変数をSurvived（１が生存、０が死亡）とし、２項ロジット分析を行った。生死を特定するあらゆるモデルを構築し、そのモデルが実際にどれほど生死を特定できたか予測用データと比較した際の正答率を調べた。そのモデルをランキングサイトのkaggleに投稿し、ランキングの上昇を試みた。

説明変数（既存）

- Pclass：部屋のグレード　1st,2nd,3rd（Crew）
- Name：乗客・クルーの名前
- Sex：性別
- Age：年齢
- SibSp：乗船している兄弟および配偶者の数
- Parch：乗船している両親および子供の数
- Ticket：チケット番号
- Fare：運賃
- Cabin：船室
- Embarked：乗船した港

説明変数の追加

fg.Mr:名前にMrがあるかどうか

fg.Mrs:名前にMrsがあるかどうか

fg.Miss:名前にMissがあるかどうか

fg.Master:名前にMasterがあるかどうか

fg.Doctor名前にDoctorがあるかどうか

Pclass2:Pclass（部屋のグレード）が2

Pclass3:Pclass（部屋のグレード）が3

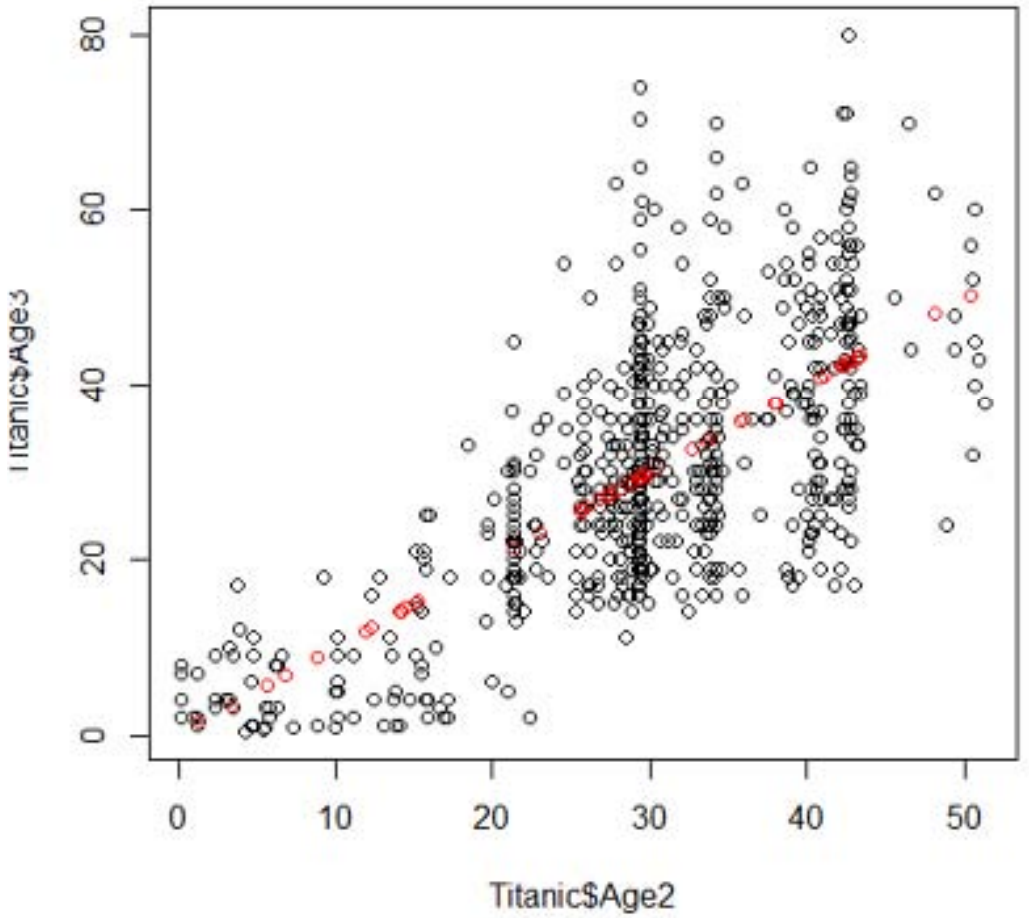
Age(年齢）の欠損値の補完
変数の追加

赤丸が補完した値

Age2 :すべて推定値

Age3 : 欠損値のみ補完し、既存の数値はそのまま

Age4: Age3 × Age3



モデルの構築①

- res1<−glm(formula = Survived ~ Sex *Pclass3 + SibSp + Age3+ fg.Mrs + fg.Miss + fg.Master + Parch, family = ″binomial″, data = Titanic)

- ヒット率　0.8406285

モデルの構築②

- res2<−glm(formula = Survived ~ Sex *Pclass3 + SibSp + Age4+ fg.Mrs + fg.Miss + fg.Master + Parch, family = ″binomial″, data = Titanic)

- ヒット率　0.8395625

- モデル①からはAge3→Age4に変更

結果

実際にkaggleのサイトにアップロードし、予測用データとのヒット率を確かめる。

モデル①・・・
５９２位（０．７７９９０）

589	new	str8nr	0.77990	5	Mon, 09 Dec 2013 16:52:24
590	new	mGTI	0.77990	12	Mon, 09 Dec 2013 19:44:56
591	new	johnrisko	0.77990	5	Mon, 09 Dec 2013 22:14:29 (-0.8h)
592	new	やみ	0.77990	1	Tue, 10 Dec 2013 09:03:02

Your Best Entry
Congratulations on making your first submission!

593	↓91	Iraquitan Cordeiro Filho	0.77512	5	Fri, 11 Oct 2013 14:27:17 (-24.7h)
594	↓91	NERV	0.77512	4	Fri, 18 Oct 2013 20:22:43 (-8d)
595	↓91	Nicholas Bayborodin	0.77512	2	Sat, 12 Oct 2013 02:19:38
596	↓91	RizOa	0.77512	4	Sat, 12 Oct 2013 10:19:24 (-0.7h)
597	↓91	William Westlin	0.77512	1	Sat, 12 Oct 2013 23:01:25

モデル②・・・
４３５位（０．７８４６９）

430	new	Mathieu Cliche	0.78469	10	Mon, 23 Dec 2013 00:12:40 (-0.1h)
431	new	WSN	0.78469	8	Mon, 23 Dec 2013 12:30:46
432	new	Manuel Lopez	0.78469	4	Mon, 23 Dec 2013 13:59:32
433	new	Kepi	0.78469	7	Mon, 23 Dec 2013 20:19:24 (-0.7h)
434	new	Mark De Blanger	0.78469	3	Mon, 23 Dec 2013 19:53:47
435	↑141	やみ	0.78469	4	Tue, 24 Dec 2013 07:49:46

Your Best Entry
You improved on your best score by 0.00478.
You just moved up 171 positions on the leaderboard.

436	↓33	Changqing Huo	0.77990	7	Tue, 05 Nov 2013 15:15:11 (-12.3d)
437	↓33	danil.valgushev	0.77990	10	Sun, 27 Oct 2013 20:49:20 (-3.2d)
438	↓33	Shipeng	0.77990	9	Sat, 26 Oct 2013 04:46:45 (-19.2h)
439	↓33	Midboss	0.77990	2	Fri, 25 Oct 2013 19:20:03
440	↓33	Kolibridrache	0.77990	4	Mon, 18 Nov 2013 19:04:36 (-23.9d)

考察

- 生存と男性であることは負の相関
- 生存と年齢には負の相関
- 生存と乗船している両親・子供の数は正の相関

“Women and children first”が実際に行われていたことがわかる。

- 生存と席の階級の低さは負の相関
- 生存と運賃の高さは正の相関

➡ 金持ち・身分の高いものが優先的に逃がされた

感想

- タイタニック号の事件は有名で、私たちも知識が多い分先入観が強く、モデルによっては相関があまりでないものもあり、試行錯誤が難しかった。

- ただ与えられたもともとの変数だけで考えるのではなく、Mr,Mrsなどの新しい変数を追加するだけでもヒット率が上がるのでとても面白みがあった。

- 自分で変数を追加するなど試行錯誤をし、実際にランキングが上がると達成感があった。

課題

- SVMモデル
- 回帰樹によるより高い因子の選出

以上を用いて、さらなるヒット率の向上が可能と考えられるので、上記をマスターし利用する。

参考文献

- http://www.titanic1.org/ (2013/12/27)
- http://www.swissinfo.ch/jpn/detail/content.html?cid=7196364 (2013/12/27)
- http://www.kaggle.com/c/titanic-gettingStarted(kaggle2013/12/24)
- http://d.hatena.ne.jp/yutakikuchi/20120827/1346024147 (R言語でSVMによる分類学習2013/12/24)