春輪講　タイタニック号の生存者予測

笹瀬研究室

61620335 吉田一輝

1 機械学習による生存予測概要

タイタニック号沈没事件の乗客データに基づいた生存予測

1. 生存情報がある(答えがある)training setを用いて
   機械学習モデルを構築する

2. test setの乗客データに対して構築した機械学習モデルで生存予測をする

予測に扱える特徴量(11種)

“PassengerId”, “PassengerClass”, “Name”, “Sex”, “Age”, “SibSp”, “Parch”, “Ticket”, “Fare”, “Cabin”, “Embarked”

生存と相関のある特徴を取捨選択する必要がある

2.1 従来方式 特徴量取捨選択

PassengerId

- 単なる通し番号であるため, 生存確率とは相関がない

Ticket

- 同じチケット番号を持つ場合, 同時に申し込みをしたことを表すので家族かそれに準ずる可能性が高い

→ それは, SibSpとParchで表されるので考慮しない

チケット番号347082の乗客(計7名)

Andersson,	Mr. Anders Johan
Andersson,	Miss. Ellis Anna Maria
Andersson,	Miss. Ingeborg Constanzia
Andersson,	Miss. Sigrid Elisabeth
Andersson,	Mrs. Anders Johan(Alfrida Konstantia Brogren)
Andersson,	Miss. Ebba Iris Alfrida
Andersson,	Master. Sigvard Harald Elias

同じSibSp + Parch=6
かつ
同じ名字を持つ
親類

2.2 従来方式 特徴量取捨選択

Sex, Age

- 1912年当時, 海難事故が発生した場合女性や子どもを先に助ける“Women and children first”という標語が採用されていた[3]ので, 生存確率に大いに関係があると予想される

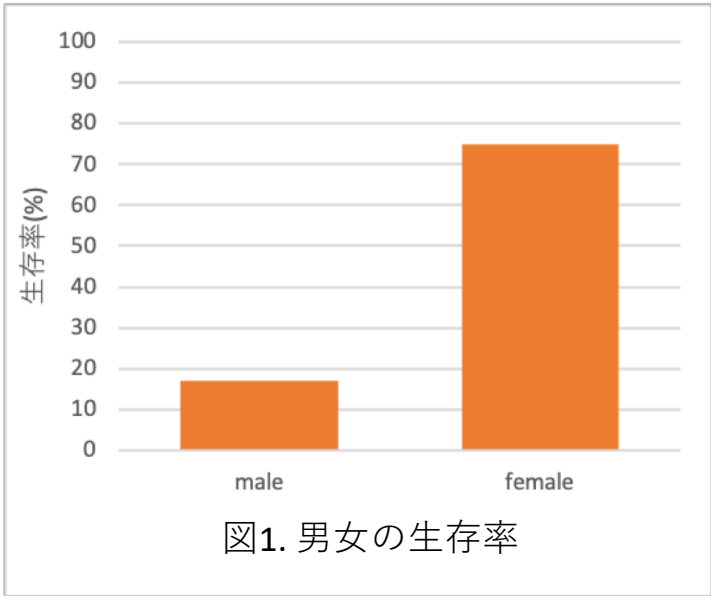


図1. 男女の生存率

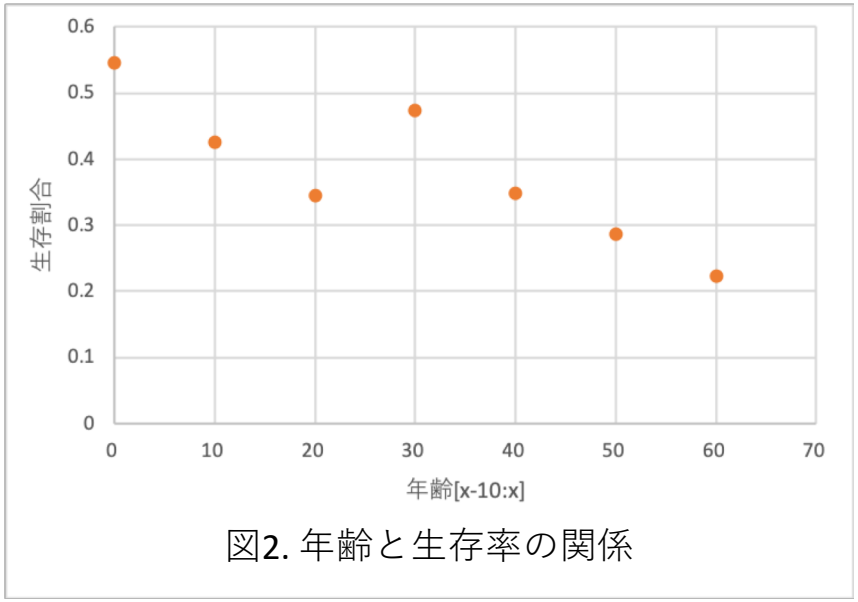


図2. 年齢と生存率の関係

- 男女で生存率に明らかな開きがあることを確認
- 年齢と生存率の相関係数は-0.864538となり, 負の相関があることを確認

[3] Mikael Elinder and Oscar Erixson. 2012. Every man for himself-Genders, Norms and Survival in Maritime Disasters.

2.3 従来方式 特徴量取捨選択

Pclass, Fare

- 明らかに階級が高い方が生存確率が高い

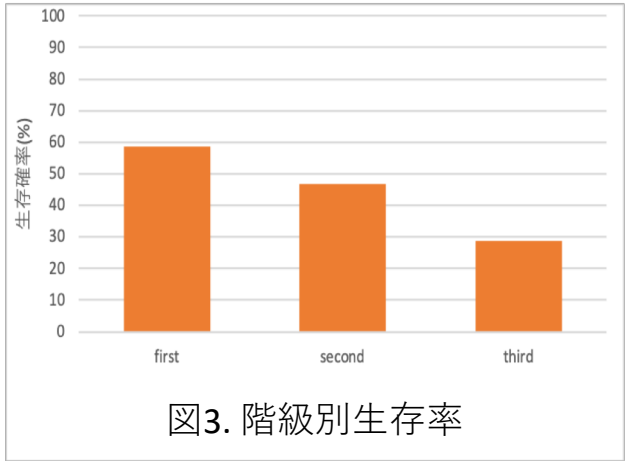


図3. 階級別生存率

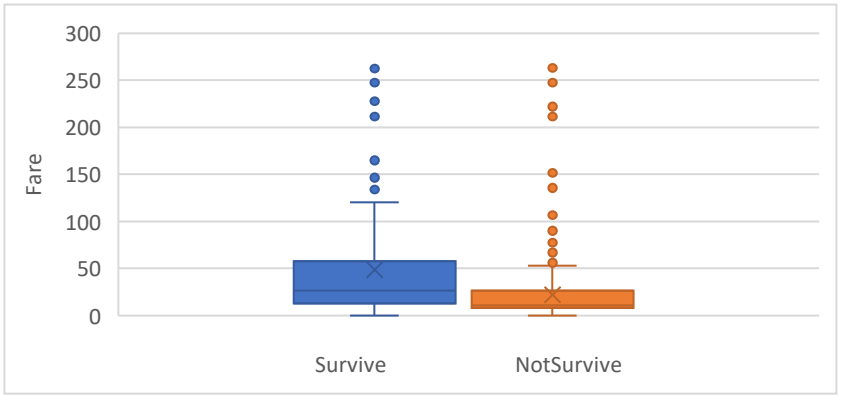


図4. 生存者と運賃の関係

SibSp, Parch

- 家族の構成人数が4人に近いほど生存確率が高くなる傾向にある

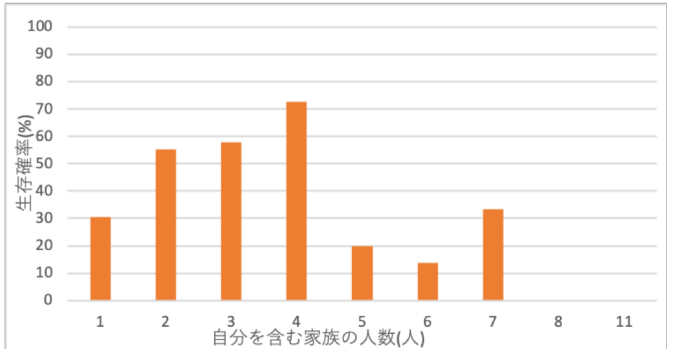


図5. 家族の人数と生存率の関係

2.4 従来方式 特徴量取捨選択

Cabin

- 客室の位置によって生存確率が異なる
- 欠損値が多いが, 欠損値の集合は明らかに死亡率が高い

欠損値は未知の部屋記号“X”とおけば高い死亡率の部屋を表すことができる

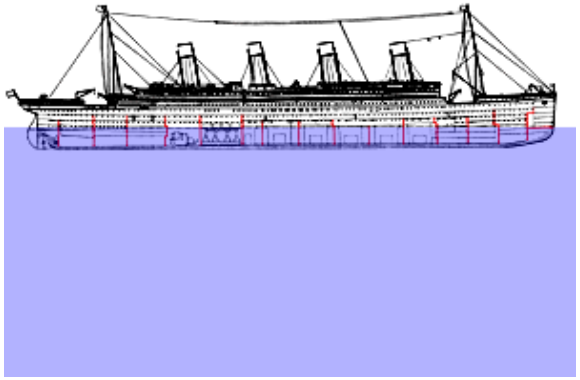


図6. タイタニック号が沈む様子[4]

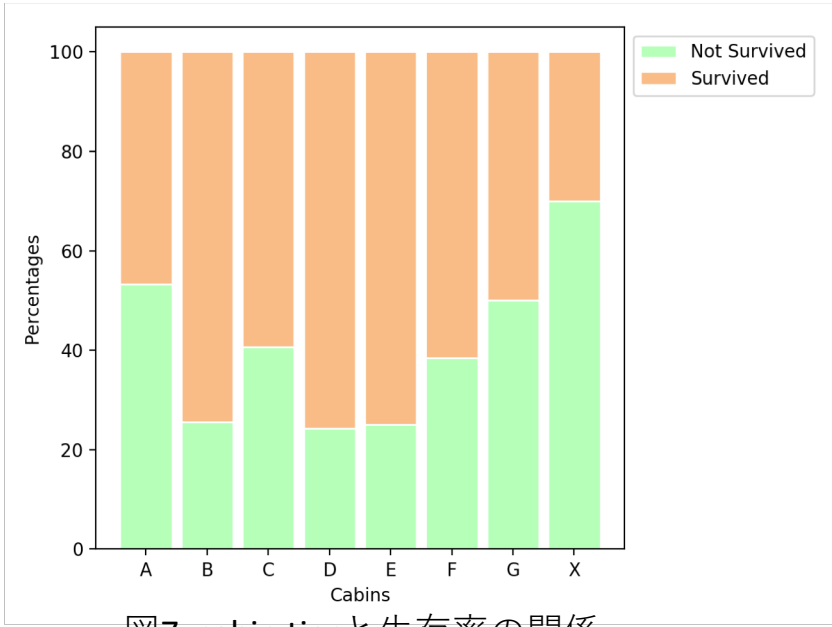


図7. cabin tierと生存率の関係

[4] File: Titanic-sinking-animation.gif. (February 19, 2012). Retrieved from https://commons.wikimedia.org/wiki/File:Titanic-sinking-animation.gif (accessed February 14, 2019)

2.5 従来方式 特徴量取捨選択

Embarked

- 乗船港C, Q, Sの順に生存率が高くなる

Name

- 同じ男性の中でもMasterやDrと敬称のついた人物は優先的に救命ボートに乗せられたと思われる

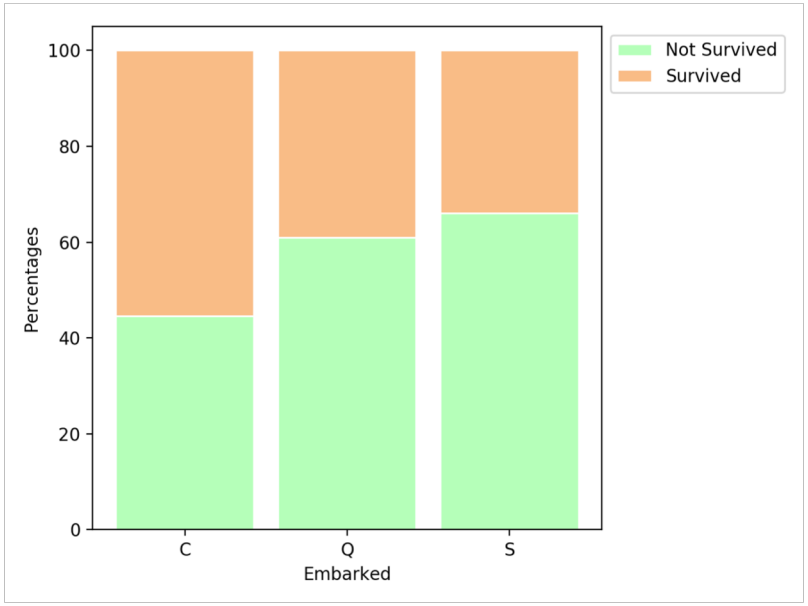


図8. 乗船港と生存確率の関係

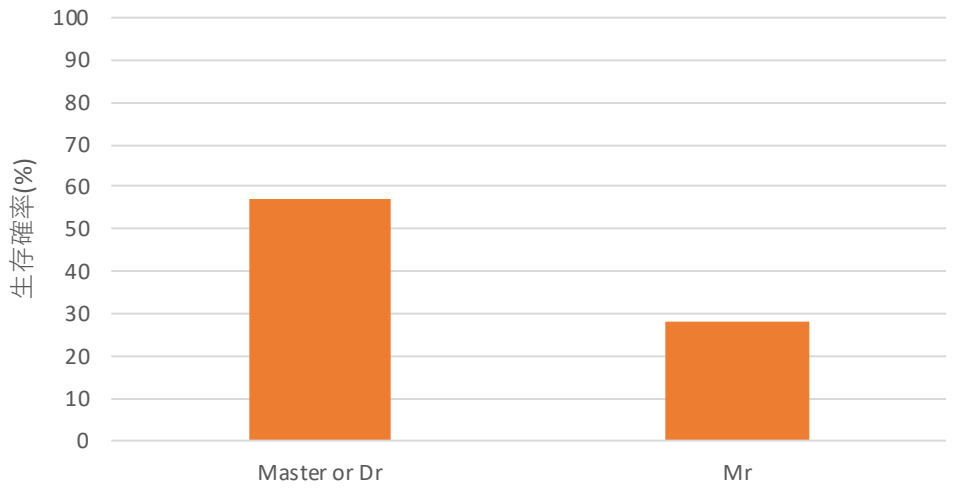


図9. 男性の敬称と生存確率の関係

2.6 従来方式 欠損値の補完

Cabin

欠損値”X”をおくことで, その他にデータに影響することなく欠損値を扱うことができる.

Embarked

2件しか欠損していないカテゴリ変数なので, 最頻値を用いる.

Fare

平均を用いると極端に運賃が高い富裕層や, 値が0の乗組員と思われる客が多大な影響を与えてしまうので中央値を用いる.

Age

年齢を妥当な判断材料として保持するため, 平均値を用いて値のもつ重みを変化させない.

2.7 従来方式 特徴量エンジニアリング

乗船している家族の人数について

- 前述の通り, 乗船している家族の人数と生存確率には相関がある. SibSpもParchも親族を表しているのでデータを一つにまとめる.

Family_Members = SibSp + Parch + 1

- また, 家族がいる場合といない場合では後者の方が生存率が若干低いので, 単身かどうかを表す特徴量を用意する.

Is_Alone = (Family_Members > 1) ? 0 : 1

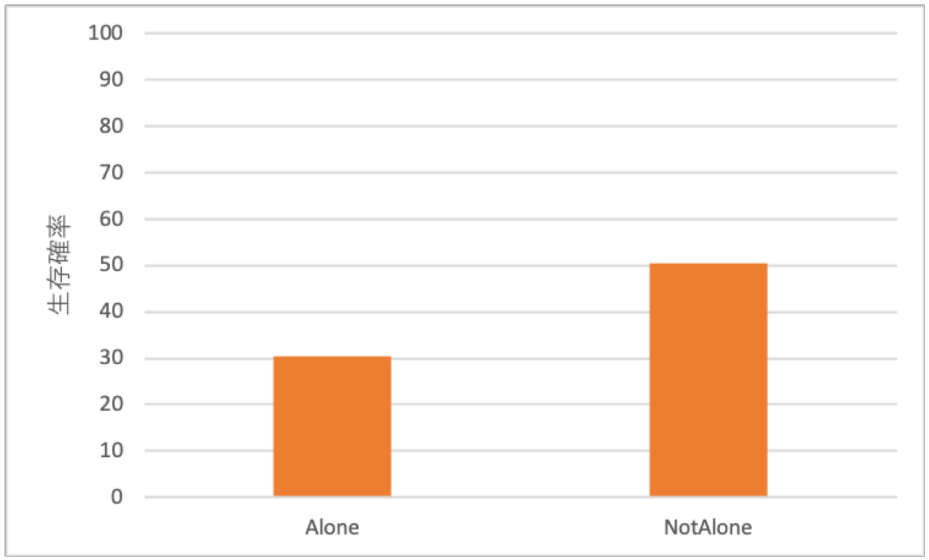


図10. 単身とそうでない人の生存率

2.8 従来方式 特徴量エンジニアリング

客室のデッキ番号について

Cabinの頭文字によって生存傾向が異なり, 数値データではないためそれぞれのフラグを立てる用の特徴量を用意する(One-Hotエンコーディング).

columns = [“Cabin_A”, “Cabin_B”, ... , “Cabin_X”]

表1. One-Hotエンコーディングの例

	Cabin A	Cabin B	Cabin C
αさん	1	0	0
βさん	0	1	0

名前について

MasterやDrの敬称のついた人物の生存率が高く, それ以外は単にユニークな文字列であるため, 敬称の部分だけを抜き出し, それぞれのフラグを立てる用の特徴量を用意する(One-Hotエンコーディング).

columns = [“Title_Master”, “Title_Miss”, ... , “Title_Mr”]

その他数値データについて

連続する値を持つ数値データの範囲が大きいので,
正規化して値の範囲を狭める

df[“Fare”] = (df[“Fare”] – df[“Fare”].mean()) / df[“Fare”].std()

3.1 従来方式 問題点

正答率は以下の表1の通りである．ただし

- family: 家族の人数を詳細に分けた場合
- family_weak: 家族の人数を1, 2〜4, 5〜人で場合分けした場合
- conventional: 従来手法を用いた場合
- cabin_drop: 従来手法からCabinでの判定をなくした場合

この4種において, 特徴量の種類が少ないほど平均が大きく, 分散が小さく, 最大値が大きくなる傾向が見られる.

- 次元の呪い(curse of dimensionality)
- One-Hotエンコーディングによる不安定性

表2. 特徴量数とスコアの関係

	平均	分散	最大値	特徴量数
cabin_drop	0.75478	2.5268E-05	0.76076	20
conventional	0.74581	3.6851E-05	0.75598	29
family_weak	0.73444	2.8632E-05	0.74162	31
family	0.73744	7.6889E-05	0.75119	36

3.2 従来方式 問題点

次元の呪い(curse of dimensionality)

- 学習データの次元が増加するにつれて, 学習に必要なサンプル数は指数関数的に増加してしまう[6]. 手元にある有限なデータでは十分な学習結果が得られず, 未知のデータに適切に対応できなくなっている.

One-Hotエンコーディングによる不安定性

- 数値データ以外を扱い, フラグを記述するために特徴量が一つ必要なOne-Hotエンコーディングは, 同じ問題に対して妥当なモデルが複数存在して係数が一意に定まらず, それが結果の解釈の妨げとなることがある[7].

[6] Michel Verleysen and Damien Francois. 2005. The Curse of Dimensionality in Data Mining and Time Series Prediction.
[7] Alice Zheng and Amanda Casari. 2019. 機械学習のための特徴量エンジニアリング. Translated by 株式会社ホクソエム, オライリージャパン, p84.

3.3 従来方式 問題点

- 家族人数をそのまま特徴量として扱っている. しかし, 単純に家族の人数が増えれば生存率が増加するといった相関は存在せず, 家族の人数に応じた生存率の傾向が現れている.

- 極端に高い運賃を払っている乗客のデータをそのまま扱っている. 平均値などの数値データがそれらの値に引っ張られてしまう.

- 年齢の欠損値を単純に平均値で補っているので, 平均年齢の乗客の人数だけが増加してしまっている. 平均値は変化しないが, 実際の年齢の分布を乱してしまっている.

4.1 提案方式

次元の呪い, One-Hotエンコーディングによる不安定性

Cabinという9種類もの特徴量を削除し, 特徴量の種類を削減することで, 機械学習の安定化

家族の人数の扱い方

まとめて扱うと相関が見出しにくい家族の人数を, 似たような値を持つ

- 一人身
- 家族の人数が2〜5のSmall Family
- 家族の人数が6人以上のLarge Family

に場合分けすることにより, より正確に家族の人数を扱う

年齢の欠損値の補完

平均値ではなく, 同じ

- Pclass
- Sex
- Title

の値を持つデータを探し, そのデータで補完することで, より現実的な欠損値の補完をする

4.2 提案方式

運賃は極端に値が大きいデータを含みばらつきが大きい

対数変換によって, 値が大きい範囲を短く狭め, 値が小さい範囲を拡大し, 運賃データをより等しい重みで扱う.

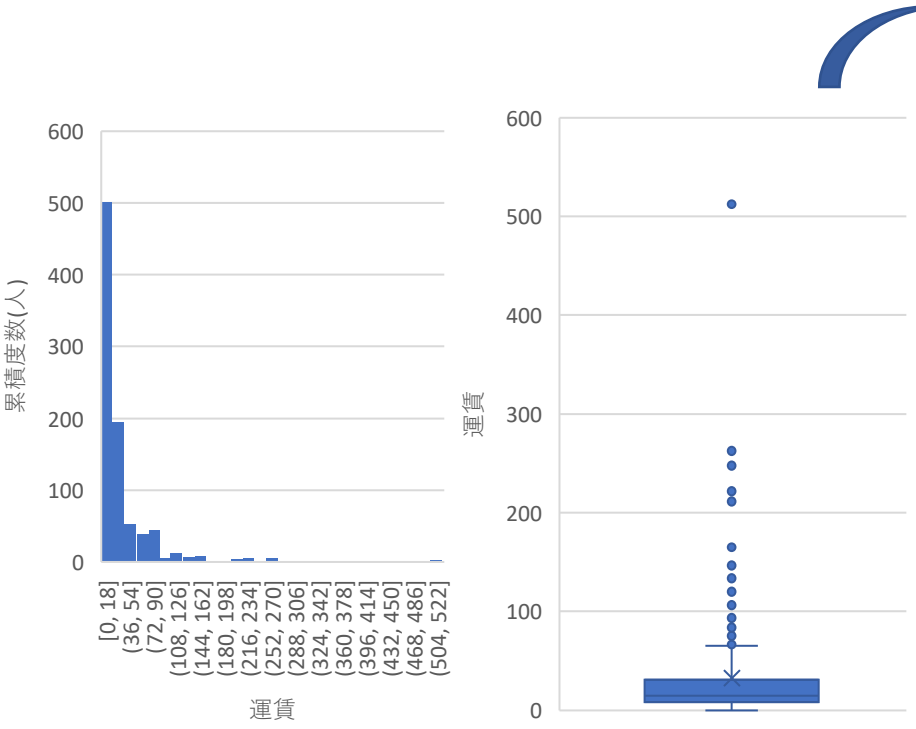


図12. 運賃とその人数

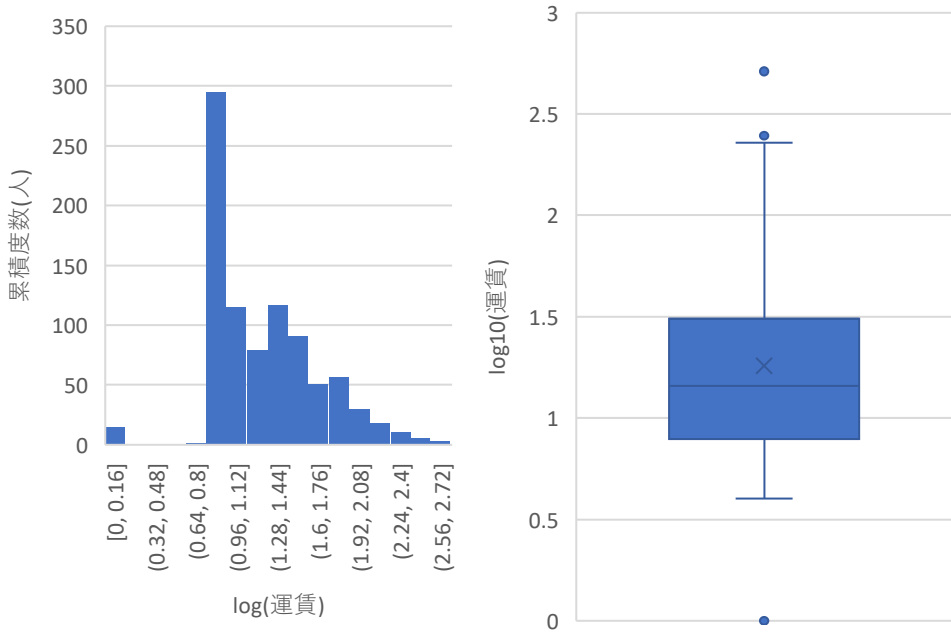


図13. log10(運賃)とその人数

4.3 提案方式 結果

- 提案方式の正答率の平均は, チュートリアルと比べて6.4%向上した
- 提案方式の正答率の最大値と最小値の差は, チュートリアルと比べて1%減少した

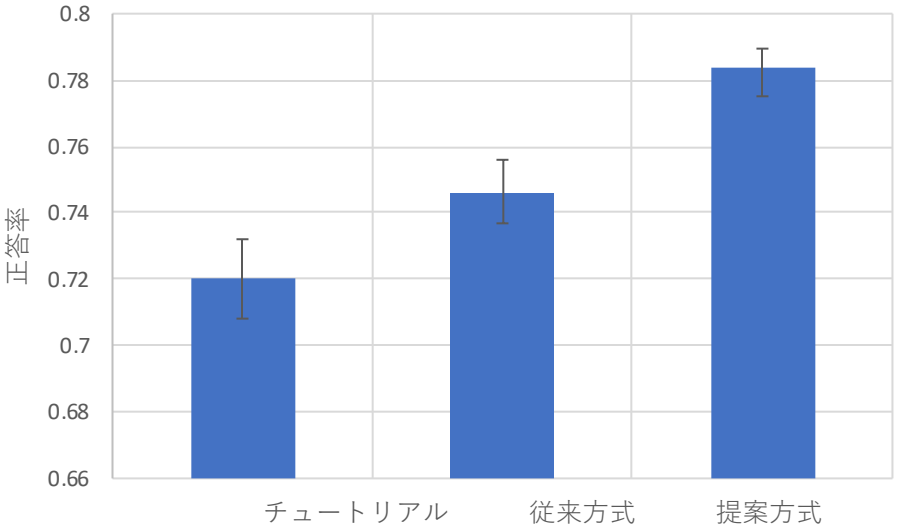


図14. 従来方式と提案方式のKaggleスコア比較

表3. 従来方式と提案方式のKaggleスコア比較

	平均	分散	最大値	特徴量数
チュートリアル[8]	0.72009	6.2965E-05	0.73205	11
従来方式	0.74581	3.6851E-05	0.75598	29
提案方式	0.78409	2.5410E-05	0.78947	23

[8] tanajun99. 機械学習によるタイタニック号の生存者予測 with Python. Retrieved from http://tanajun99.hatenablog.com/entry/2015/06/24/020007 (accessed March 5, 2019)

4.4 提案方式 考察

特徴量の種類を削減

指数関数的に増加する特徴量の組み合わせ数を削減し, 分散の減少(スコアの安定)とスコアの上昇を図れた.

家族の人数の場合分け

正負の相関と一概に言えない特徴量から, より正確な場合分けにより, スコアの上昇を図れた.

年齢の欠損値の予測補完

階級, 性別, 敬称が同じデータを探し出すことで, 年齢ヒストグラムをほとんど変化させずに欠損値を補完し, スコアの上昇を図れた.

運賃の対数変換

数値のとる幅に対してほとんどの値が小さすぎるデータ分布を, 大きい値ほど小さくなる対数変換により改善し, スコアの上昇を図れた.

まとめ

課題概要

- タイタニック号事件の乗客データから, 機械学習により生存者を予想
- 簡単のため, 木数100のdefault Random Forest Algorithmを使用

従来手法

- チケット番号, 乗客IDなどの生存と関係のない特徴量を削除
- 家族の人数や敬称などの特徴量をつくる(特徴量エンジニアリング)

従来手法の問題点

- 特徴量の種類が多い
- 家族の人数の扱い方が適切でない
- 年齢を平均値で補完
- 運賃のデータが極端に高いものと低いものに引っ張られている

提案手法

- Cabinという9種類もの特徴量数が必要なものを削除
- 家族の人数を3通りに場合分け
- 年齢を, 階級と性別と敬称から予測補完
- 運賃を対数変換し, データの偏りを和らげる

参考文献

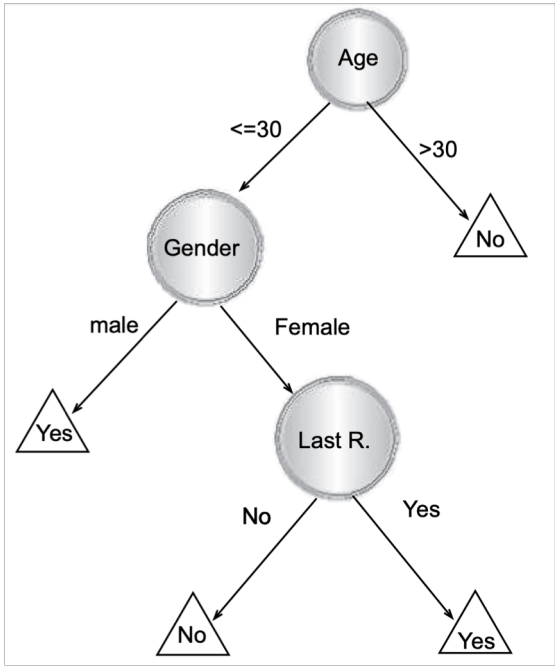
[1] Lior Rokach and Oded Maimon. 2005. *Data Mining and Knowledge Discovery Handbook,* Springer, p166.
[2] Leo Breiman. 2001. Random Forests. *Machine Learning,* 45(1), 5-32.
[3] Mikael Elinder and Oscar Erixson. 2012. Every man for himself-Genders, Norms and Survival in Maritime Disasters.
[4] File: Titanic-sinking-animation.gif. (February 19, 2012). Retrieved from https://commons.wikimedia.org/wiki/File:Titanic-sinking-animation.gif (accessed February 14, 2019)
[5] Güneş Evitan. Titanic Survival NN Approach. Retrieved from https://www.kaggle.com/gunesevitan/titanic-survival-nn-approach (accessed February 15, 2019)
[6] Michel Verleysen and Damien Francois. 2005. The Curse of Dimensionality in Data Mining and Time Series Prediction.
[7] Alice Zheng and Amanda Casari. 2019. 機械学習のための特徴量エンジニアリング. Translated by 株式会社ホクソエム, オライリージャパン, p84.
[8] tanajun99. 機械学習によるタイタニック号の生存者予測 with Python. Retrieved from http://tanajun99.hatenablog.com/entry/2015/06/24/020007 (accessed March 5, 2019)

Appendix. シュミュレーション諸元

今回の課題の目的は, 特徴量の取捨選択によるデータ分析であり, 再現性が重要である

⇒ default Random Forestに限定しても一般性を失わない

Random Forest Algorithmとは[2]



- 分散が等しく互いに独立なランダムベクトルを決定木の生成に用いる
- 各データを決定木ごとに根から葉まで条件分岐させることにより分類する
- 各決定木で出した答えで多数決を取り, 一番もっともらしいクラスに分類する
- この過程はscikit-learnが担う

図14. Decision Tree Presenting Response to Direct Mailing[1]