Titantic Problemset

Name: Philipp Neurauter (Matrikel Nr. 12018515) & Lukas Neurauter (Matrikel Nr. 12143372)

EN

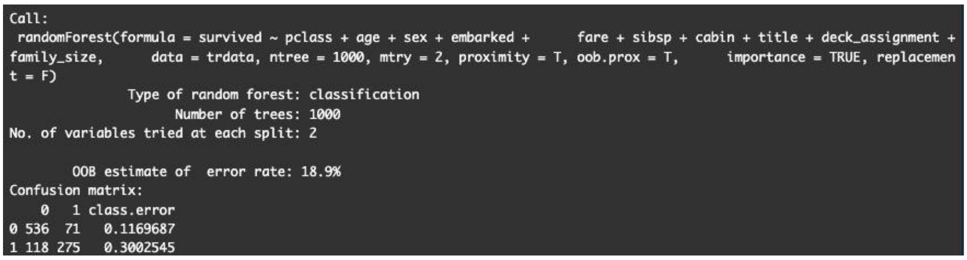
Feature engineering

By using feature engineering, additional parameters were defined and used for analysis.

- Cab -> here the values were split and only the first letter of the cabin was used
- Deck -> Additionally, the deck was assigned using the first letters. 4 categories have been added from the cabins:
    - Upper deck
    - Middle deck
    - Under deck
    - Unknown
- Family size -> This size was determined based on the last name of the passengers.
- Titles -> were also split and assigned via the name. Provides information about marriage, family, etc.
- Boats -> can also be used additionally. Two lines with an indecision were found here. Parameter quickly becomes trivial, as you get 98% accuracy OOS, close to 0% error rate. Already implies our prediction.

  Note: Line 167 & Line 78 (data)

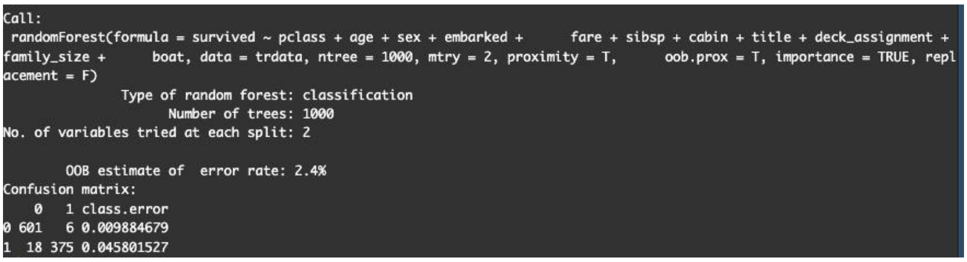
RF<-randomForest(survived ~ pclass + age + gender + embarked + fare + SIBSP + cabin + title + deck_assignment + family size, data = trdata, ntree=1000, mtry= 2, proximity=T, oob.prox=T, importance =TRUE, Substitute=F)



```
Call:
 randomForest(formula = survived ~ pclass + age + sex + embarked +      fare + sibsp + cabin + title + deck_assignment +
family_size,      data = trdata, ntree = 1000, mtry = 2, proximity = T, oob.prox = T,      importance = TRUE, replacemen
t = F)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.9%
Confusion matrix:
    0   1 class.error
0 536  71   0.1169687
1 118 275   0.3002545
```

RF<-randomForest(survived ~ pclass + age + gender + embarked + fare + SIBSP + cabin + title + deck_assignment + family size + boat, data = trdata, ntree=1000, mtry= 2, proximity=T, oob.prox=T , importance=TRUE, replacement=F)



```
Call:
 randomForest(formula = survived ~ pclass + age + sex + embarked +      fare + sibsp + cabin + title + deck_assignment +
family_size +      boat, data = trdata, ntree = 1000, mtry = 2, proximity = T,      oob.prox = T, importance = TRUE, repl
acement = F)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 2.4%
Confusion matrix:
    0   1 class.error
0 601   6 0.009884679
1  18 375 0.045801527
```

Data set cleansing / model adjustment or matching

For example, a standard can be stored for the cabins (where were the most people accommodated - where did they stay). Empty values for age -> use mean or median.

Also previous R code in the file.

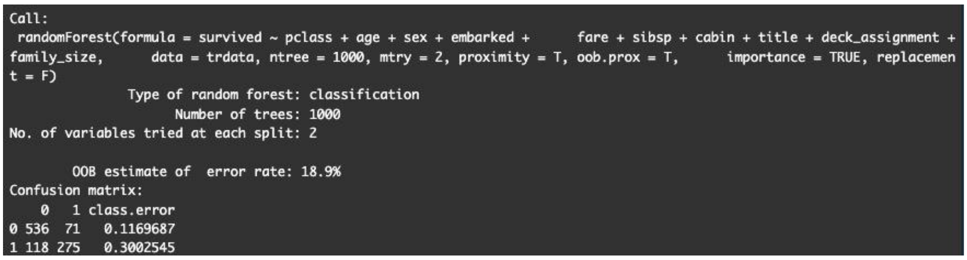
DE

Feature Engineering

Durch Einsatz von Feature Engineering wurden weitere Parameter definiert und zur Analyse herangezogen.

- Cabin -> hier wurden die Werte gesplittet und nur der Anfangsbuchstabe der Kabine verwendet
- Deck -> zusätzlich wurde durch den Anfangsbuchstaben eine Zuordnung des Deckes gemacht. Es wurden 4 Kategorien aus den Kabinen extrahiert:
    - Oberdeck
    - Mittleres Deck
    - Unterdeck
    - Unbekannt
- Family Size -> diese Größe wurde anhand des Nachnames der Passagiere ermittelt.
- Titel -> wurden auch via den Namen gesplittet und zugeordnet. Liefert Informationen zu Ehe, Familie etc.
- Boats -> kann auch zusätzlich herangezogen werden. Hier wurden zwei Zeilen mit einer Unschlüssigkeit gefunden. Parameter ist fast schon trivial, da man eine 98% Genauigkeit OOS erhält, nahe zu 0% Fehlerquote. Impliziert bereits schon unsere Vorhersage.
  Notiz: Zeile 167 & Zeile 78 (Data)

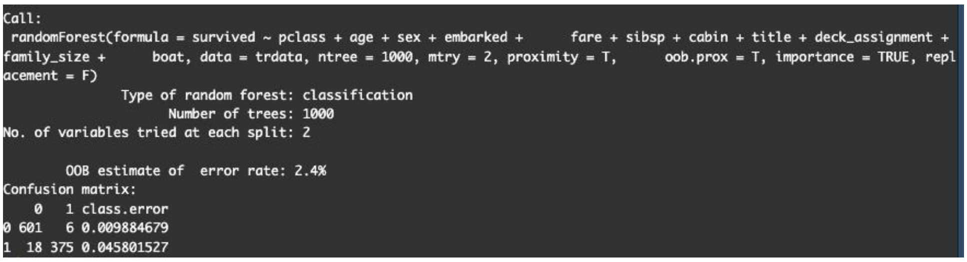
RF<-randomForest(survived ~ pclass + age + sex + embarked + fare + sibsp + cabin + title + deck_assignment + family_size, data = trdata, ntree=1000, mtry= 2, proximity=T, oob.prox=T, importance=TRUE, replacement=F)



```
Call:
 randomForest(formula = survived ~ pclass + age + sex + embarked +      fare + sibsp + cabin + title + deck_assignment +
family_size,      data = trdata, ntree = 1000, mtry = 2, proximity = T, oob.prox = T,      importance = TRUE, replacemen
t = F)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.9%
Confusion matrix:
    0   1 class.error
0 536  71   0.1169687
1 118 275   0.3002545
```

RF<-randomForest(survived ~ pclass + age + sex + embarked + fare + sibsp + cabin + title + deck_assignment + family_size + boat, data = trdata, ntree=1000, mtry= 2, proximity=T, oob.prox=T,   importance=TRUE, replacement=F)



```
Call:
 randomForest(formula = survived ~ pclass + age + sex + embarked +      fare + sibsp + cabin + title + deck_assignment +
family_size +      boat, data = trdata, ntree = 1000, mtry = 2, proximity = T,      oob.prox = T, importance = TRUE, repl
acement = F)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 2.4%
Confusion matrix:
    0   1 class.error
0 601   6 0.009884679
1  18 375 0.045801527
```

Dataset Bereinigung / Model Anpassung bzw. Fitting

Zum Beispiel bei den Kabinen kann noch ein Standard hinterlegt werden (wo waren die meisten Leute untergebracht – wo haben diese sich aufgehalten). Leere Werte bei Alter -> Einsatz von Mittelwert oder Median.

Zudem vorhergehender R-Code in der Datei.