Congressional voting

Hubert Ruczyński, Bartosz Sawicki

One hot encoding!

- Łatwiejszy
- Nie trzeba pamiętać kodowania
- Użycie w Pipeline
- Zawsze oryginalna ramka
- Lepsze accuracy

```
val = {'n':-1, '?':0, 'y':1}
party = {'democrat':2, 'republican':-2}

for column in df_num.columns:
  df_num[column] = df_num[column].map(val)

df_party = df_party.map(party)

df_num = pd.concat([df_num,df_party], axis=1)
```

```
one_hot_encoder = ce.OneHotEncoder()

one_hot = one_hot_encoder.fit_transform(X,y)
```

Selekcja cech

Nasze korzyści:

- Brak szumu w danych
- Nieco lepsze wyniki

Potencjalnie:

- *Szybsze trenowanie i działanie modelu*
- *Tańsze przechowywanie danych*

Najlepsze modele

Gradient Boosting Classifier

- Parametry:
  - Max_depth = 3
  - N_estimators = 50
- Wyniki:
  - Accuracy (5 fold CV): 97%

XGBoost

- Parametry
  - Learning_rate = 0.025
  - Max_depth = 4
  - N_estimators = 400
- Wyniki:
  - Accuracy(5 fold CV): 97%

- Taka sama macierz pomyłek
- Błędy predykcji w obserwacjach, które miały physician_fee_freeze różne od większości swojej grupy

Gradient Boosting Classifier

- Początkowo słabe wyniki
- Tuning hiperparametrów
- Najlepszy model

GridSearchCV

- Trwa długo
- Nocne poszukiwania
- Alternatywne podejście
  - Randomized Parameter Optimization
  - HalvingGridSearchCV *(eksperymentalny)*
  - Algorytmy genetyczne (brak w sklearn)

- Współpraca online
- Kompatybilny z .ipynb
- Cichy komputer
- Większa moc obliczeniowa

- Utrudniony dostęp do plików
- Problemy z zapisem przy jednoczesnej pracy
- Wygasająca sesja