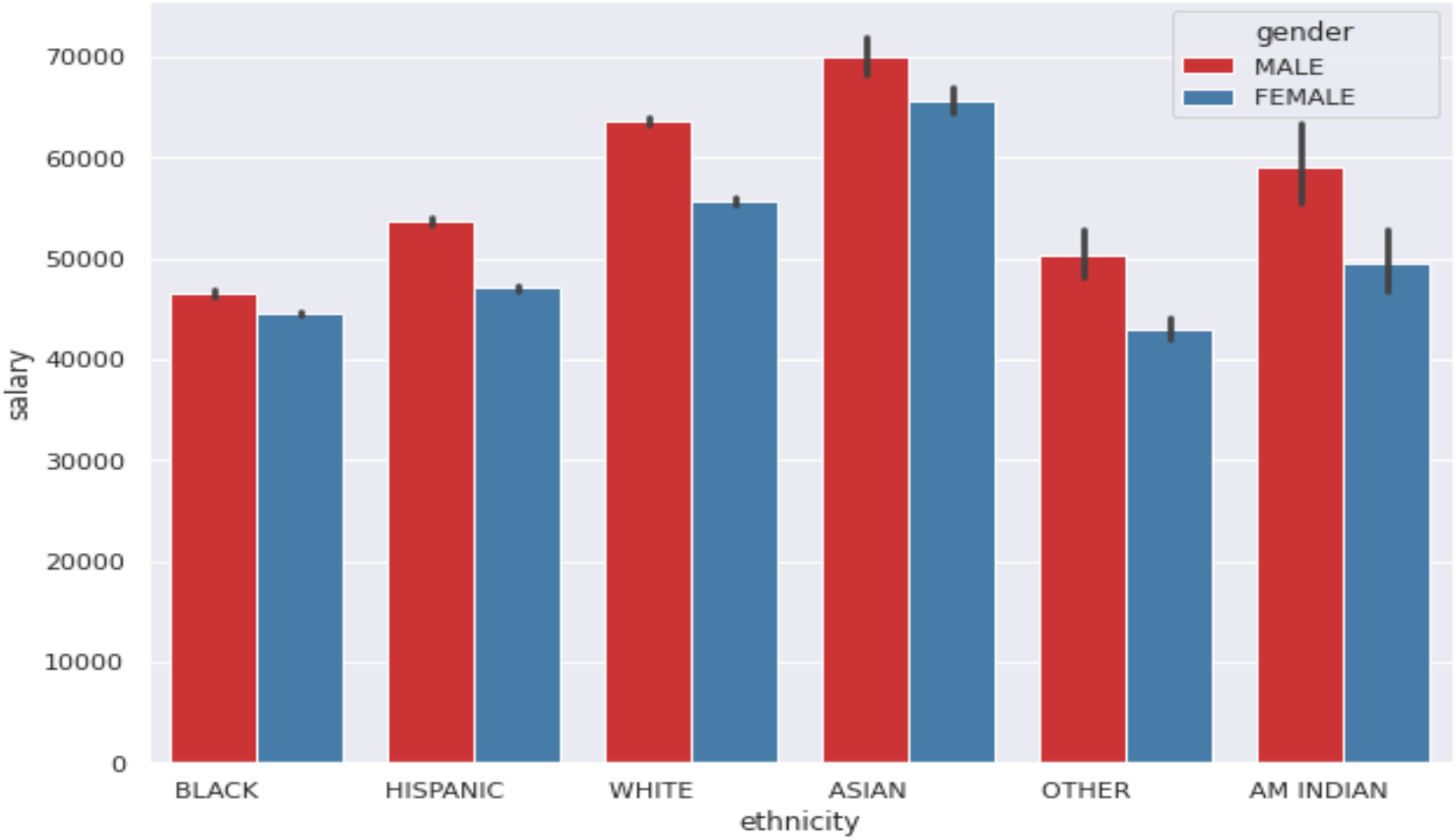
Data Challenge Presentation

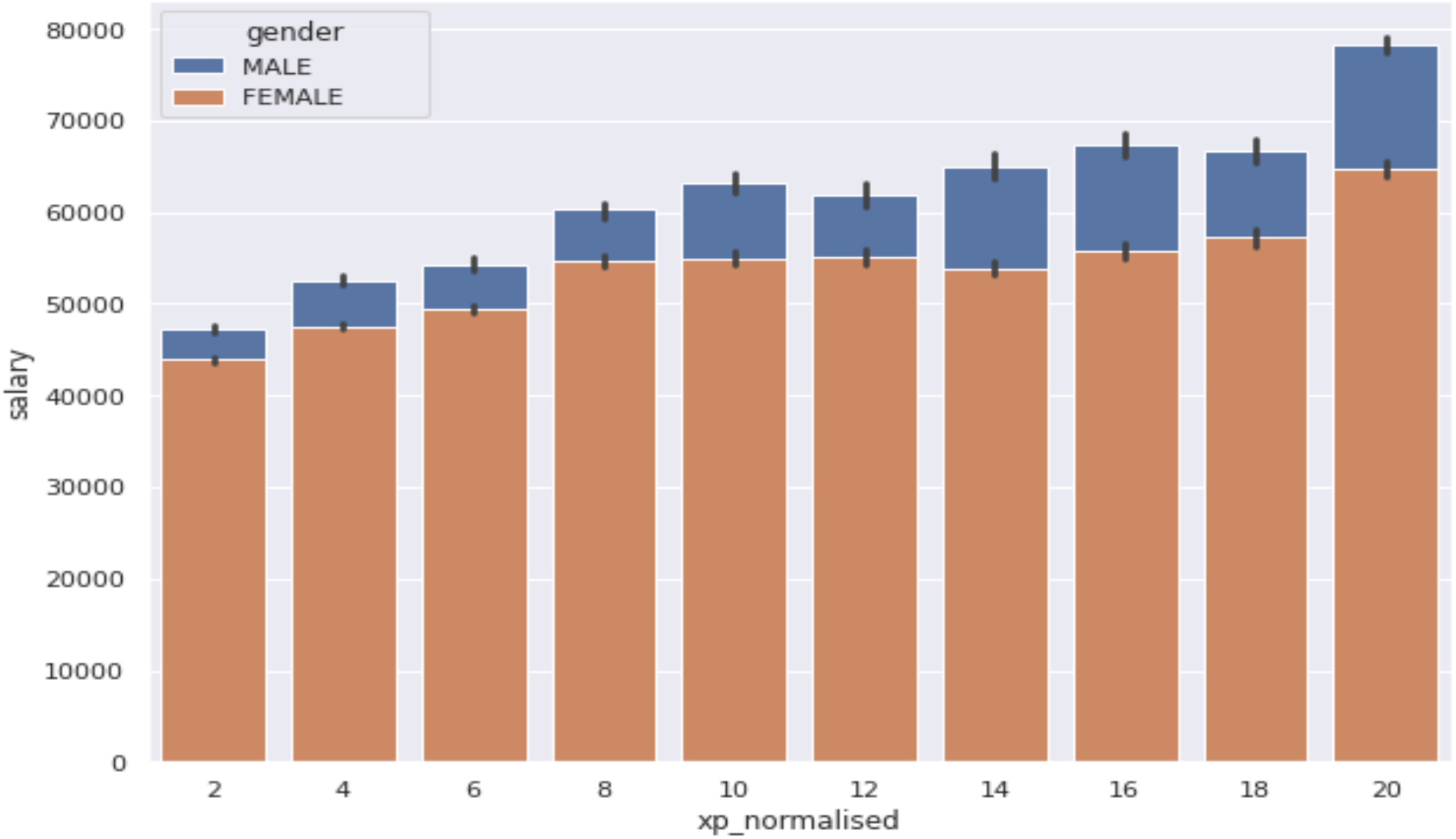
Jean Barré

April 4, 2022

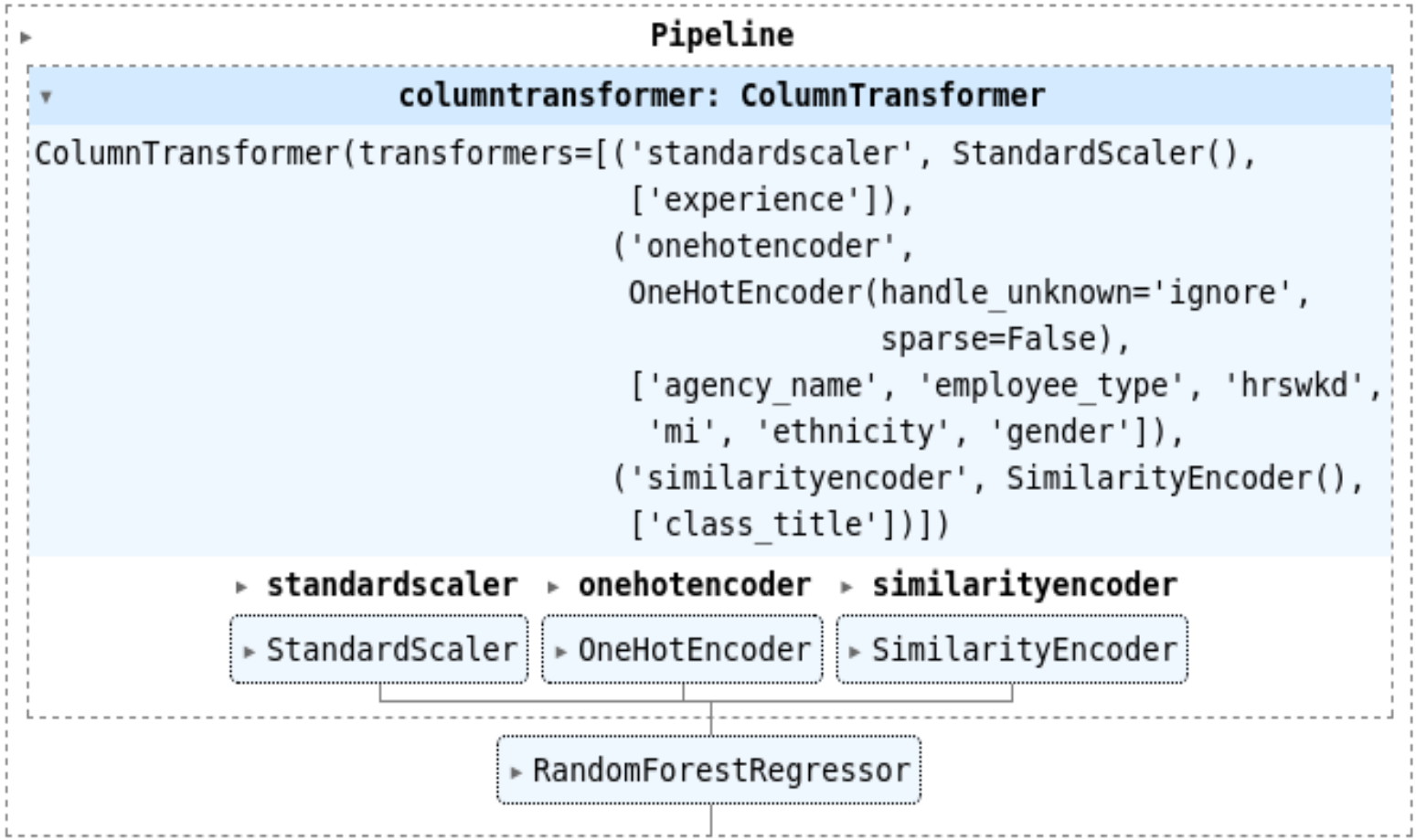
Data Visualisation 1/2



Data Visualisation 2/2



Pipeline


Pipeline
columntransformer: ColumnTransformer
ColumnTransformer(transformers=[('standardscaler', StandardScaler(),
['experience']),
('onehotencoder',
OneHotEncoder(handle_unknown='ignore',
sparse=False),
['agency_name', 'employee_type', 'hrswkd',
'mi', 'ethnicity', 'gender']),
('similarityencoder', SimilarityEncoder(),
['class_title'])])
standardscaler
onehotencoder
similarityencoder
StandardScaler
OneHotEncoder
SimilarityEncoder
RandomForestRegressor

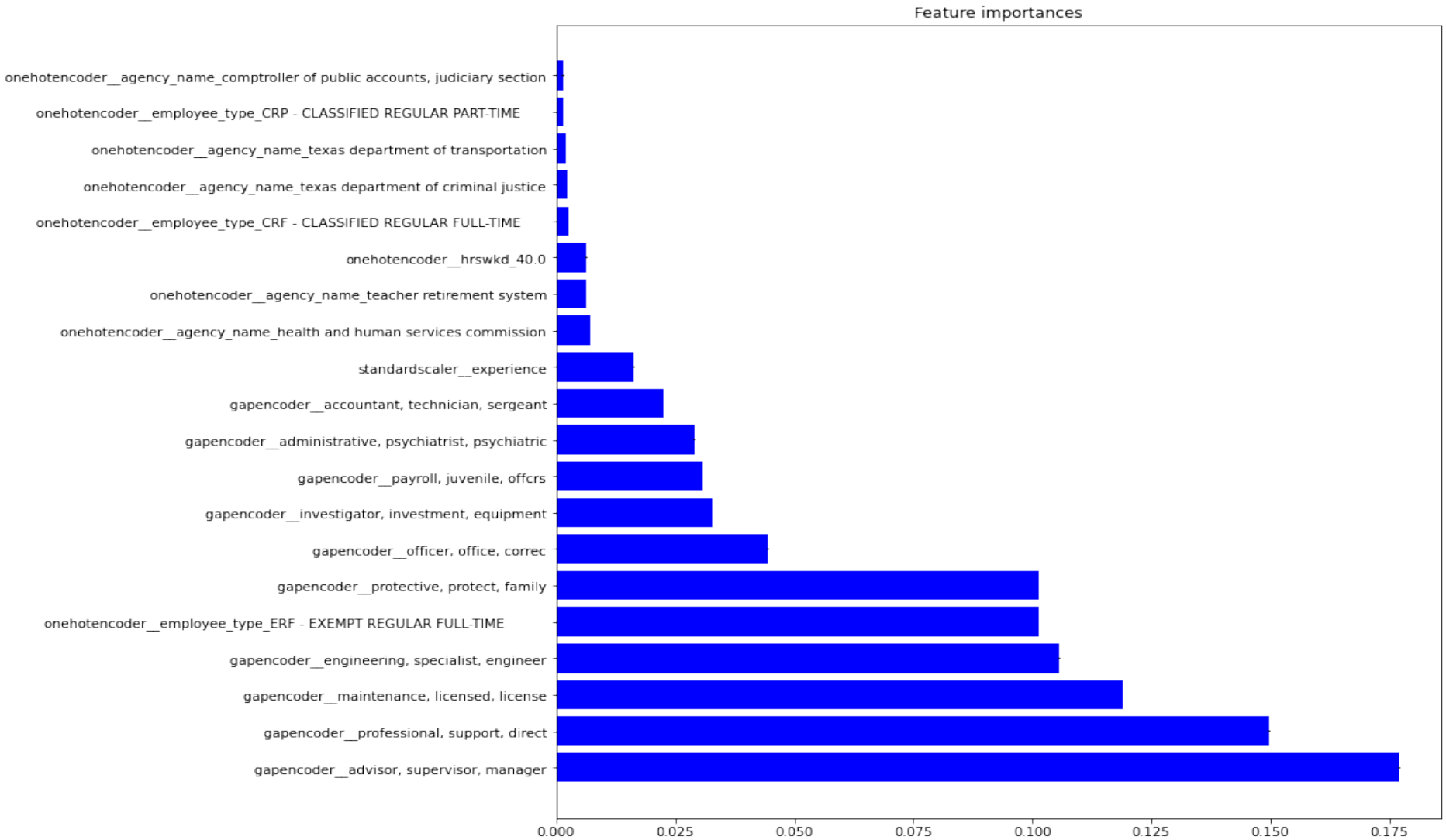
Model Evaluation

Best predictions

Model Pipeline	R2 score	MAE	RMSE
HGB & OHE	83.1%	$7203	$12207
HGB & GE	88.8%	$5203	$9207
HGB & SE	**94.6%**	**$3534**	**$6363**
RF & OHE	87.6%	$4047	$9685
RF & GE	91.2%	$3614	$8128
RF & SE	**95.0%**	**$2944**	**$6189**

Table: Results evaluation

Feature importance visualisation



Conclusion

1. Scikitlearn tools for predictive data analysis
2. Understand model inferences
    1. encoding strategies matter
    2. close look on feature importance
3. Bonus. Effects of gender on salaries.

Code available at
`https://github.com/crazyjeannot/data_challenge`