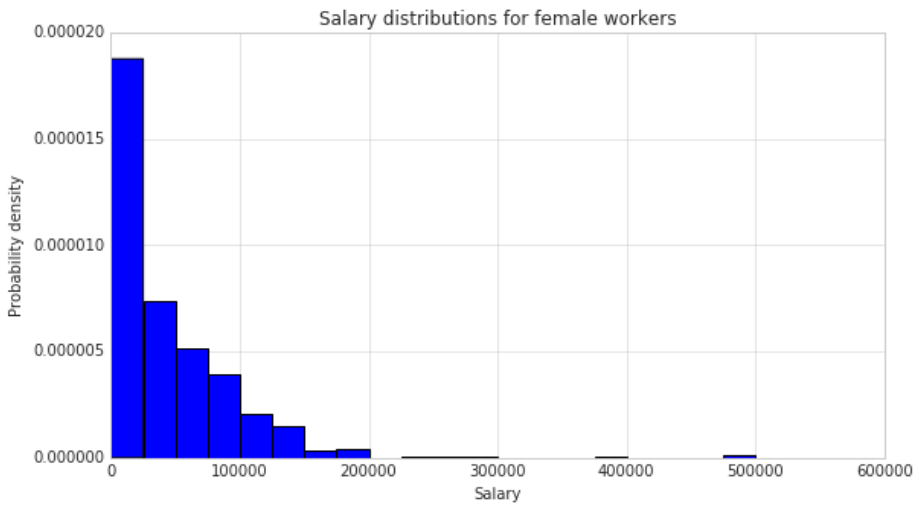
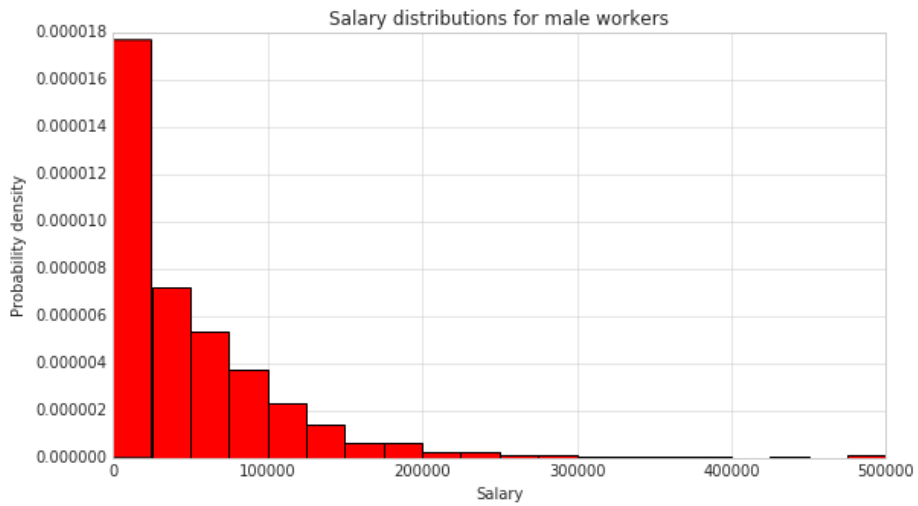
Salary Predictions Study Summary

Tiago Fernandes Lins

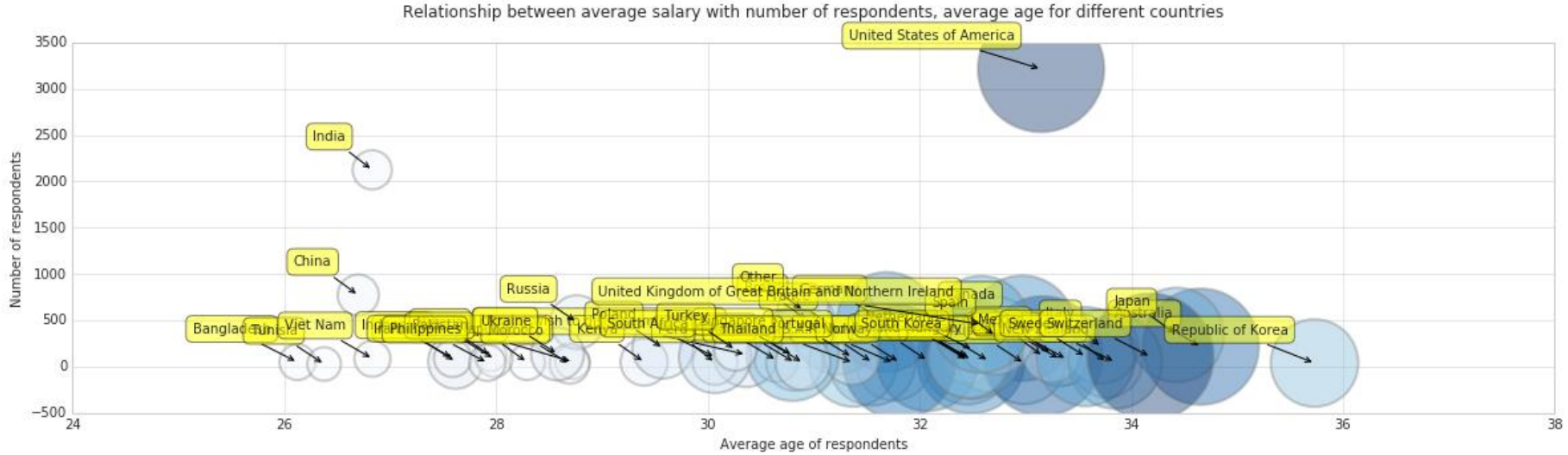
Data Visualization

- Salary distribution among males and females appear similar, but there tends to be more high outliers for males



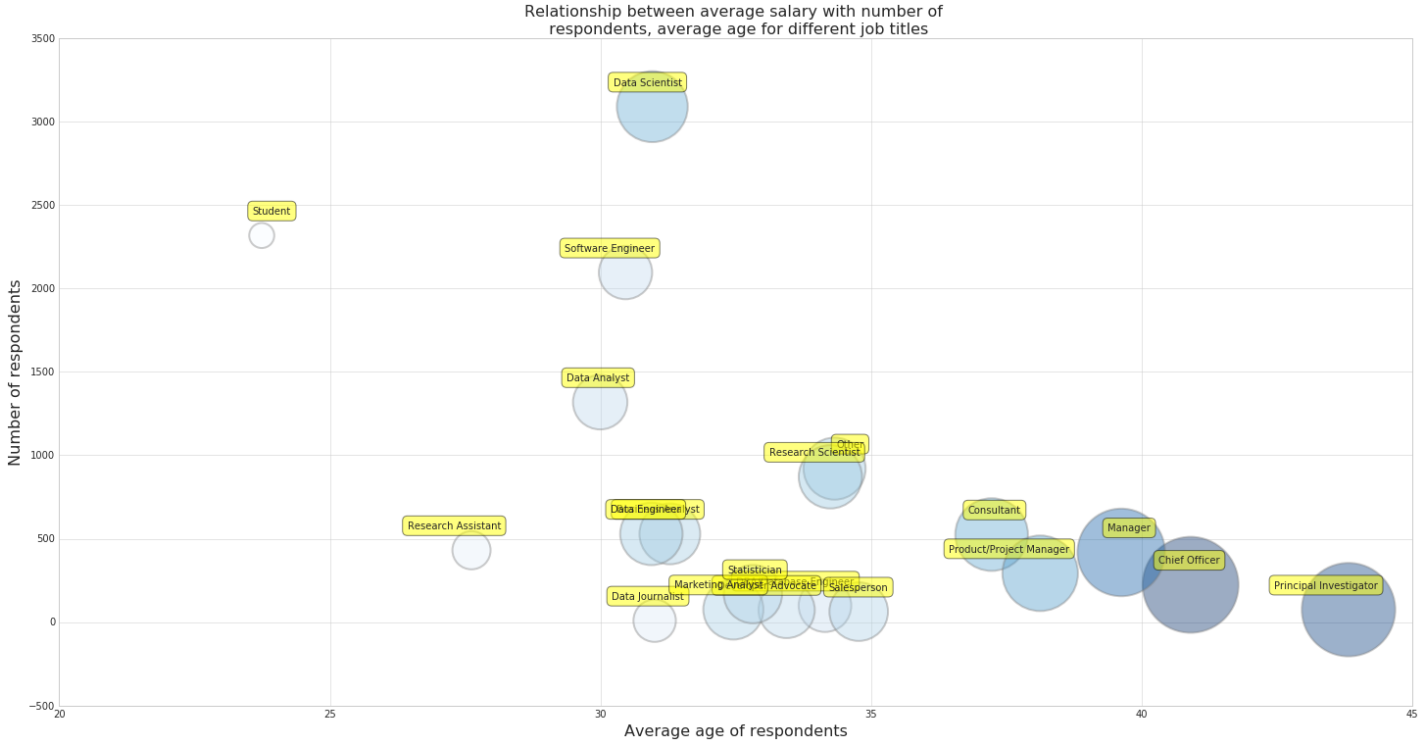


- Data points from United states and India appear to have a more observable effect on salary

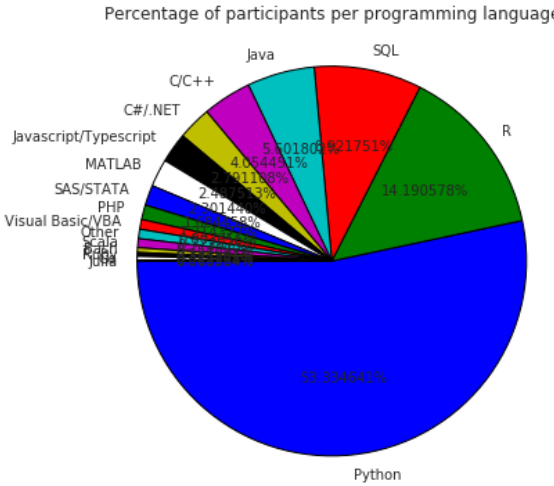


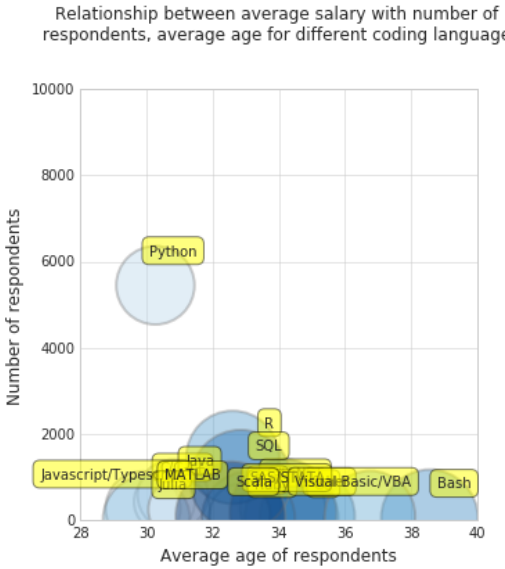
Data visualization

- Interesting trends for tools, techniques and experience for data scientists

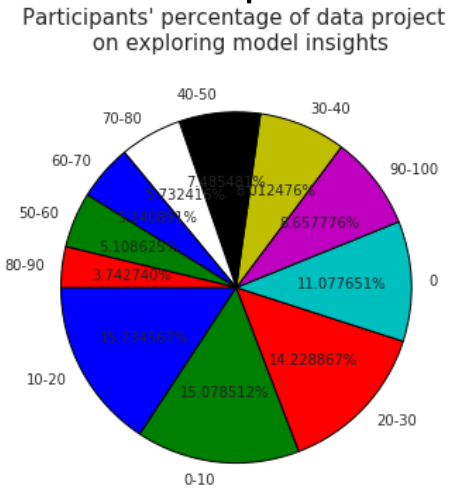


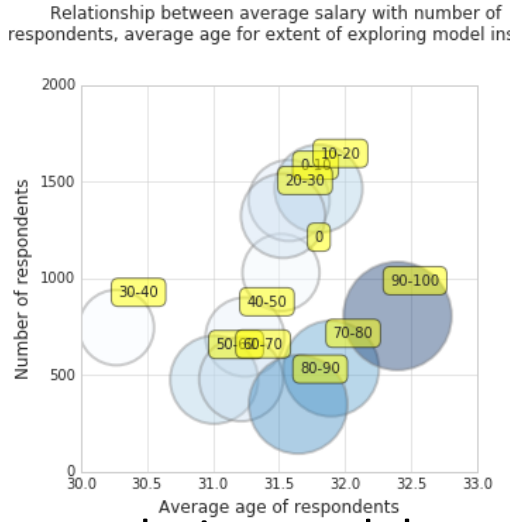
Different career pathways based on age and experience





Python most commonly used, but no obvious relationship with salary





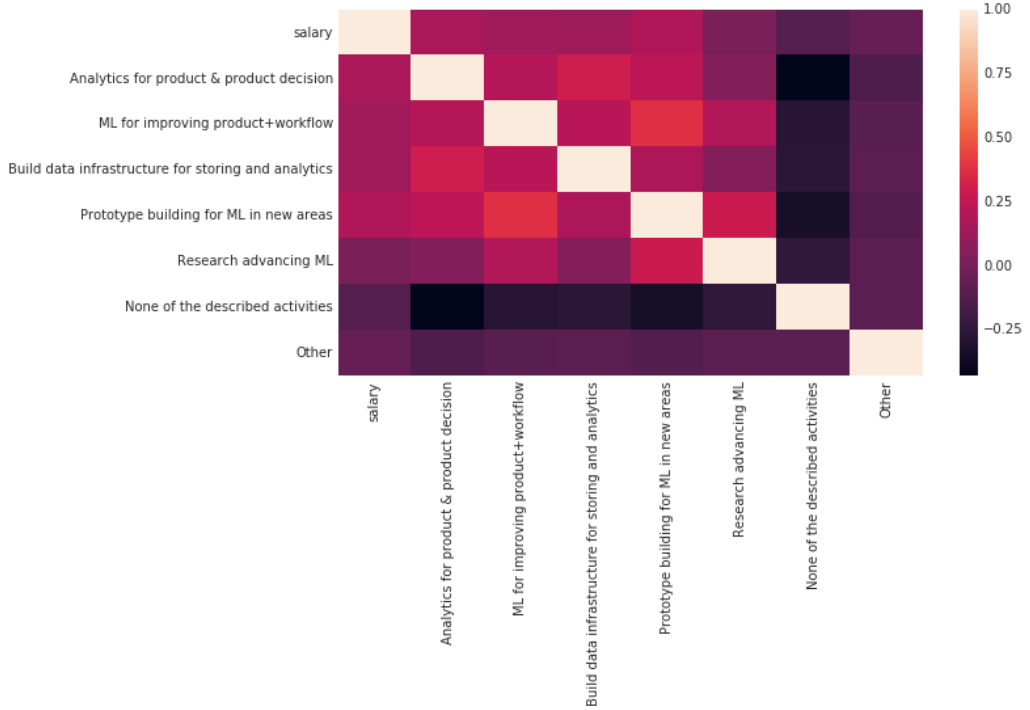
High percentage of time exploring model insights related to higher median salary

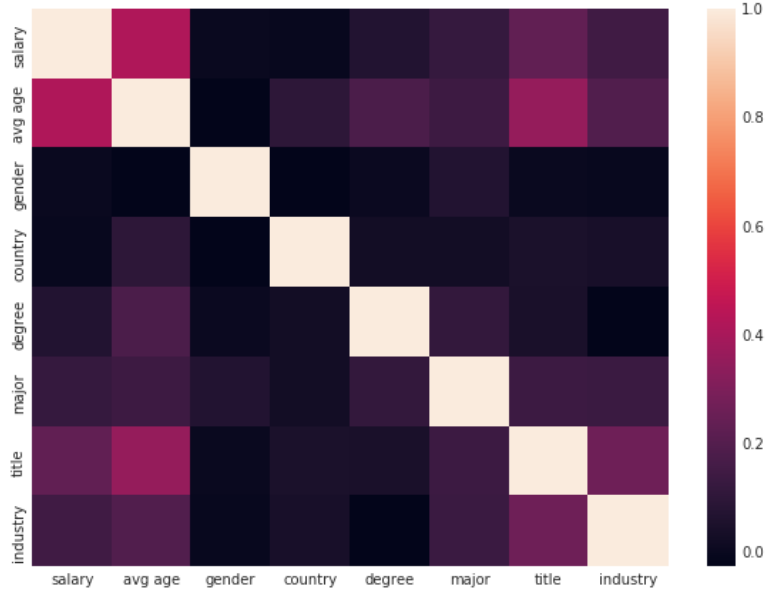
Feature Importance

Evaluated through multiple tests: correlation plots, F-test with salary data, LASSO and decision trees

Feature importance	Features
16076.78	country_United States of America
6913.296	avg age
6706.105	role_experience
5460.469	title_Chief Officer
4236.274	ML_incorporation
3412.384	country_Switzerland
3308.995	ML_years
3024.864	title_Manager
3001.495	industry_Accounting/Finance
2917.384	country_Australia

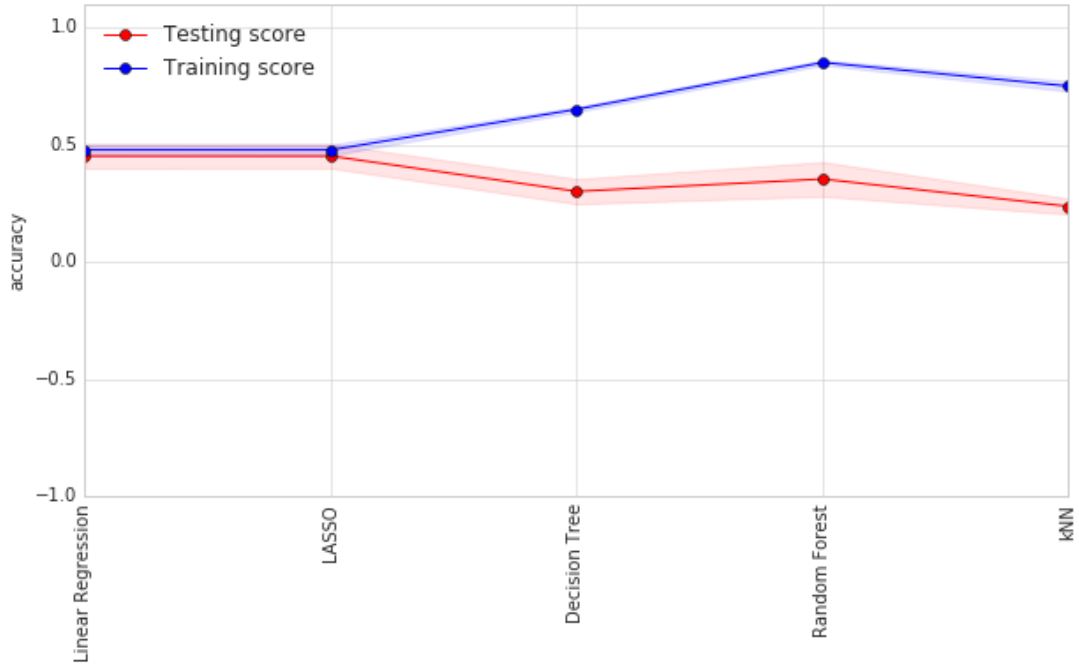
Top 10 feature by LASSO with their respective regression coefficient

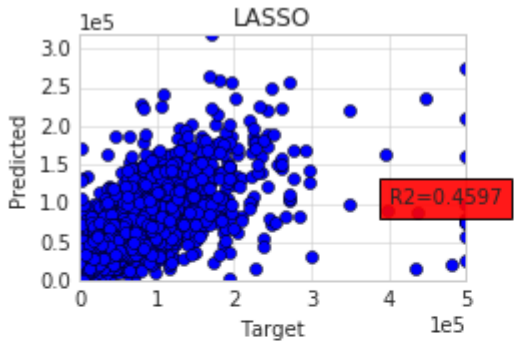


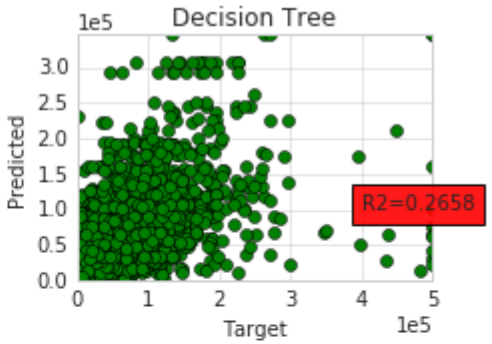


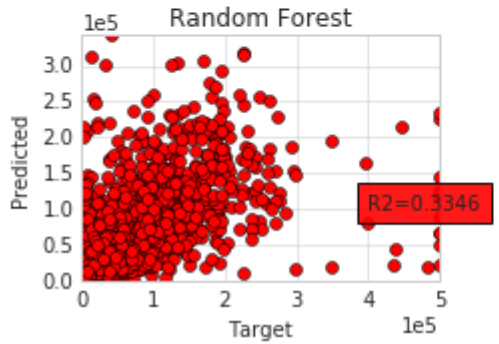
Training regression algorithms

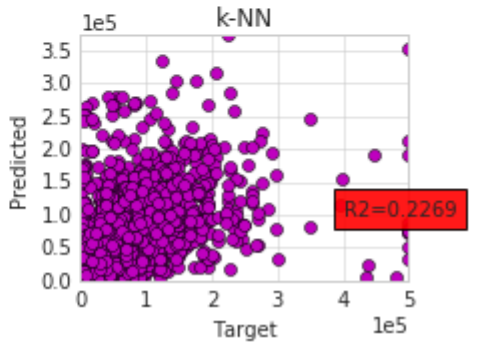
- Univariate feature selection combined with feature engineering $(x_1x_2, x_1x_3, \ldots, x_{n-1}x_n)$ are included and used for training





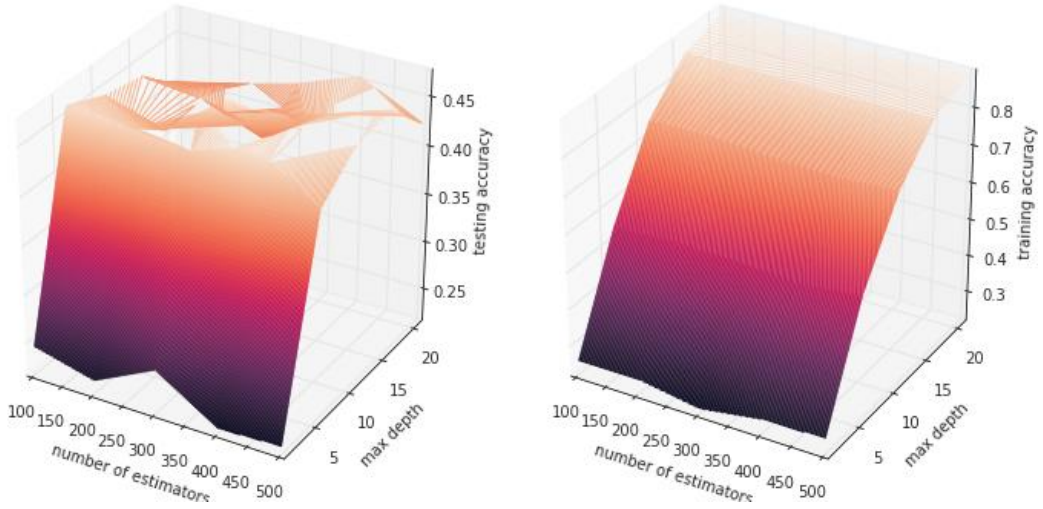






Training regression algorithms

- Hyperparameter tuning
- Slight reduction of overfit, R^2 still low generally
- K-NN showed lowest score overall, Random Forest showed highest training score



Optimized neural network