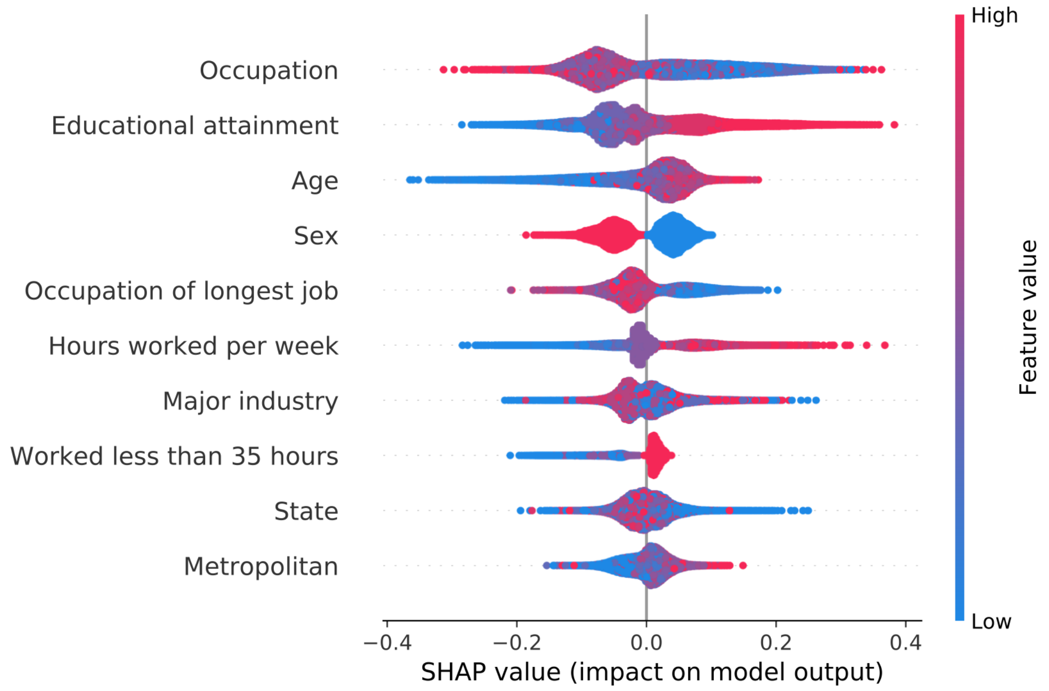
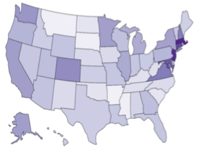
What affects our salary. Analysis and predictions based on the 2007 CPS



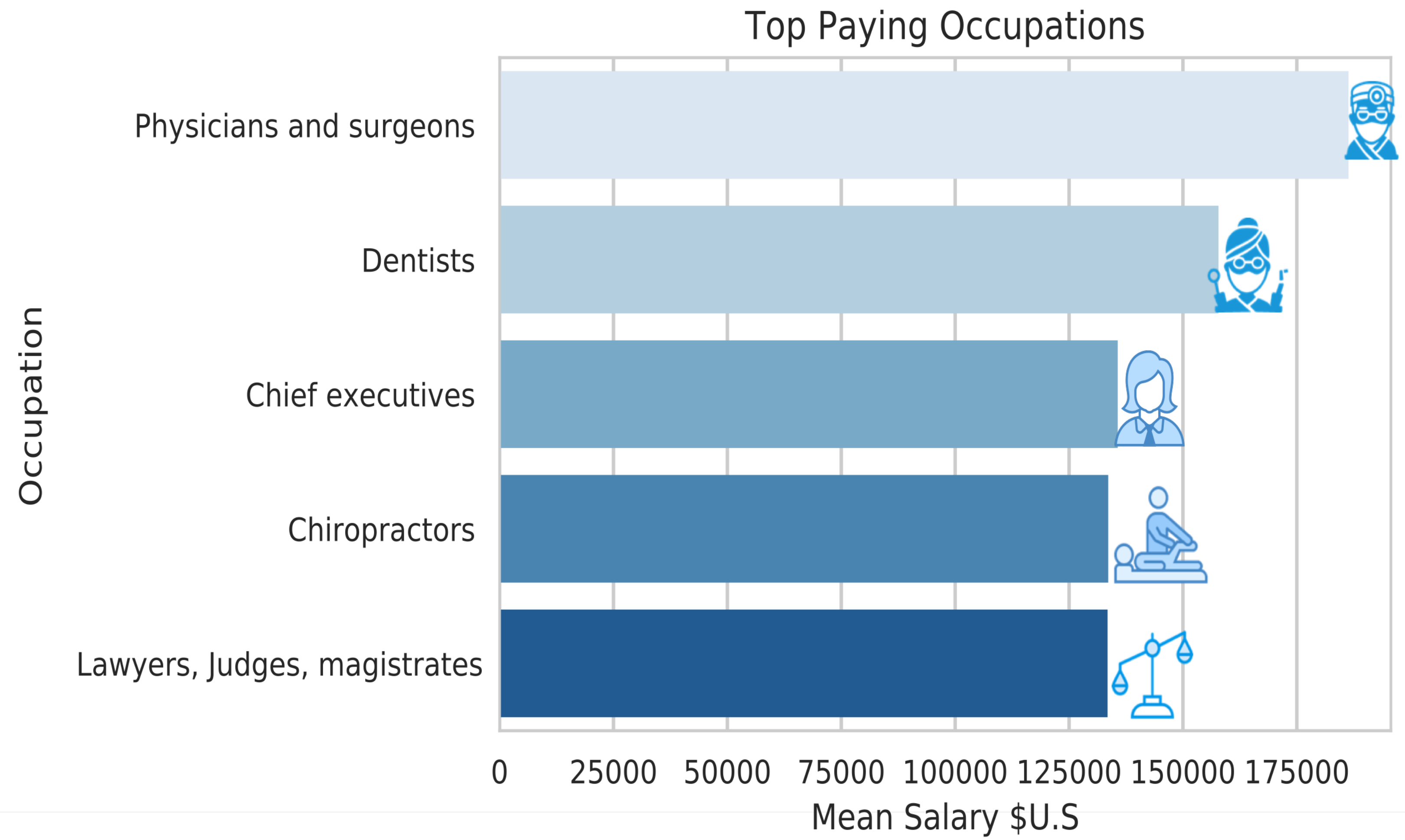
Project Goals

- Statistical analysis of the *2007 United States Department of Commerce Current Population Survey*
- *Identifying importance features which impact salary (for example: Occupation, age, gender etc.)*
- *Creating a classifier which to identify individuals which earns less then $40K per year*



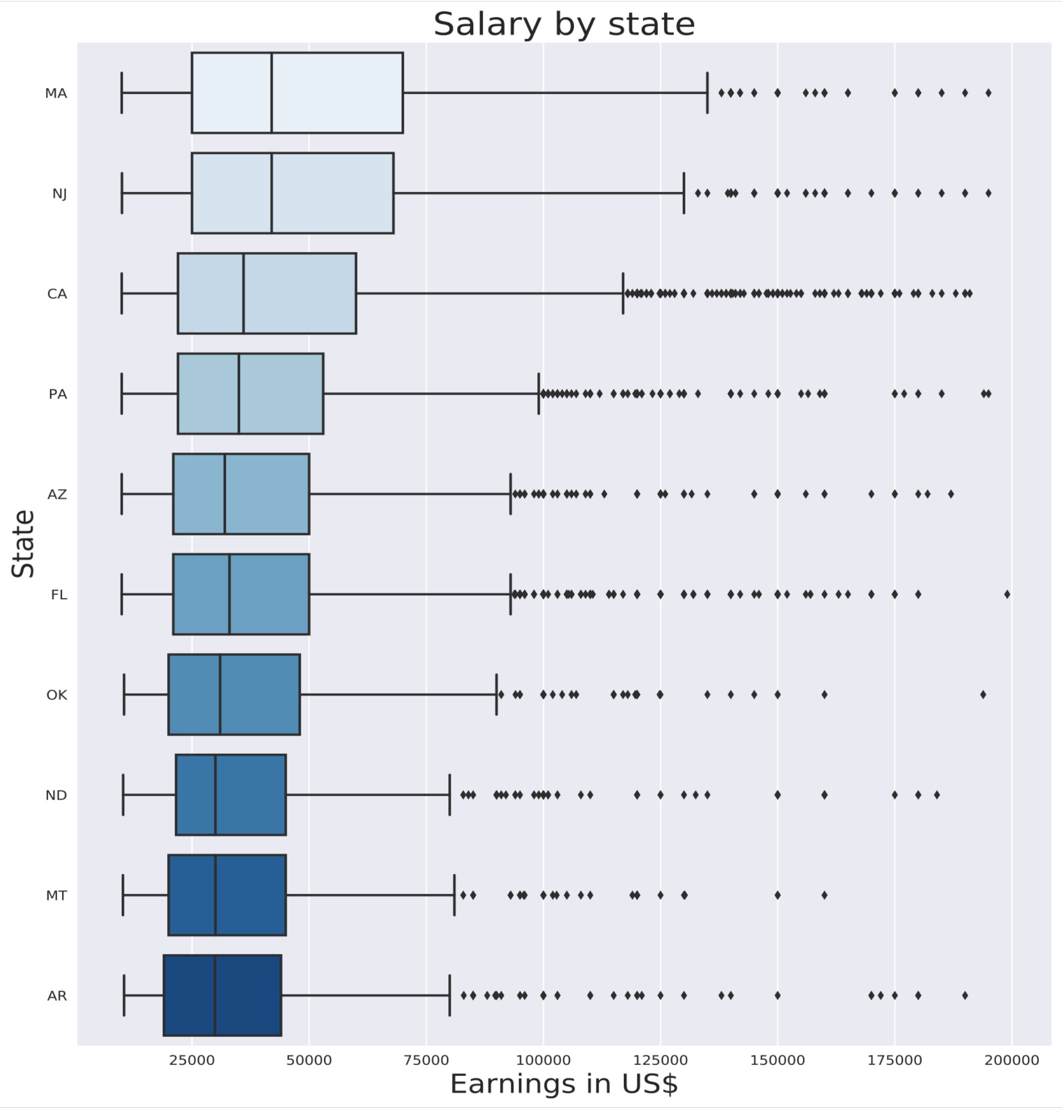
What affects our salary?

Occupation



What affects our salary?

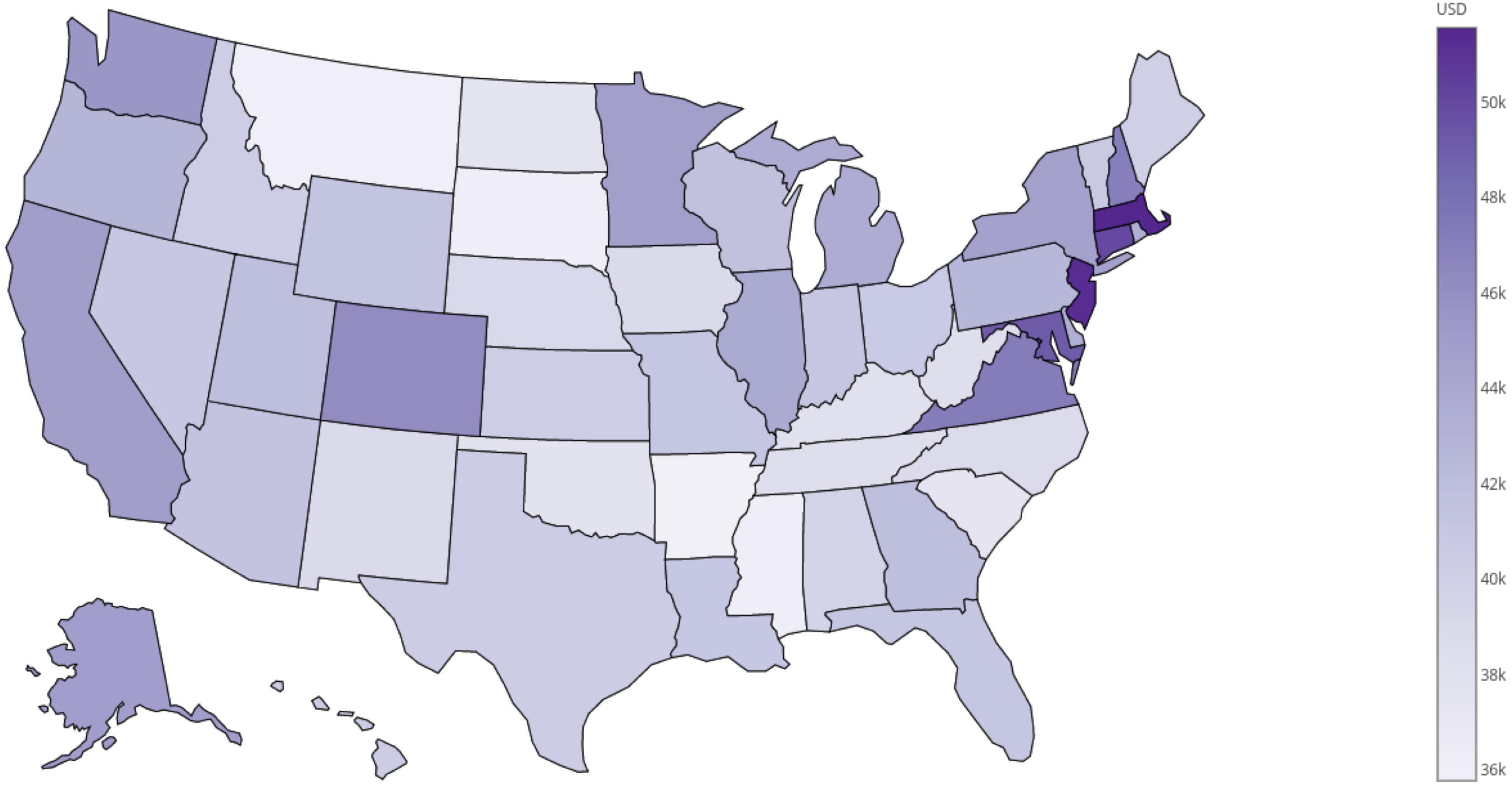
State



- Different states have different mean earnings

- *There is also a difference in the outliers distribution*

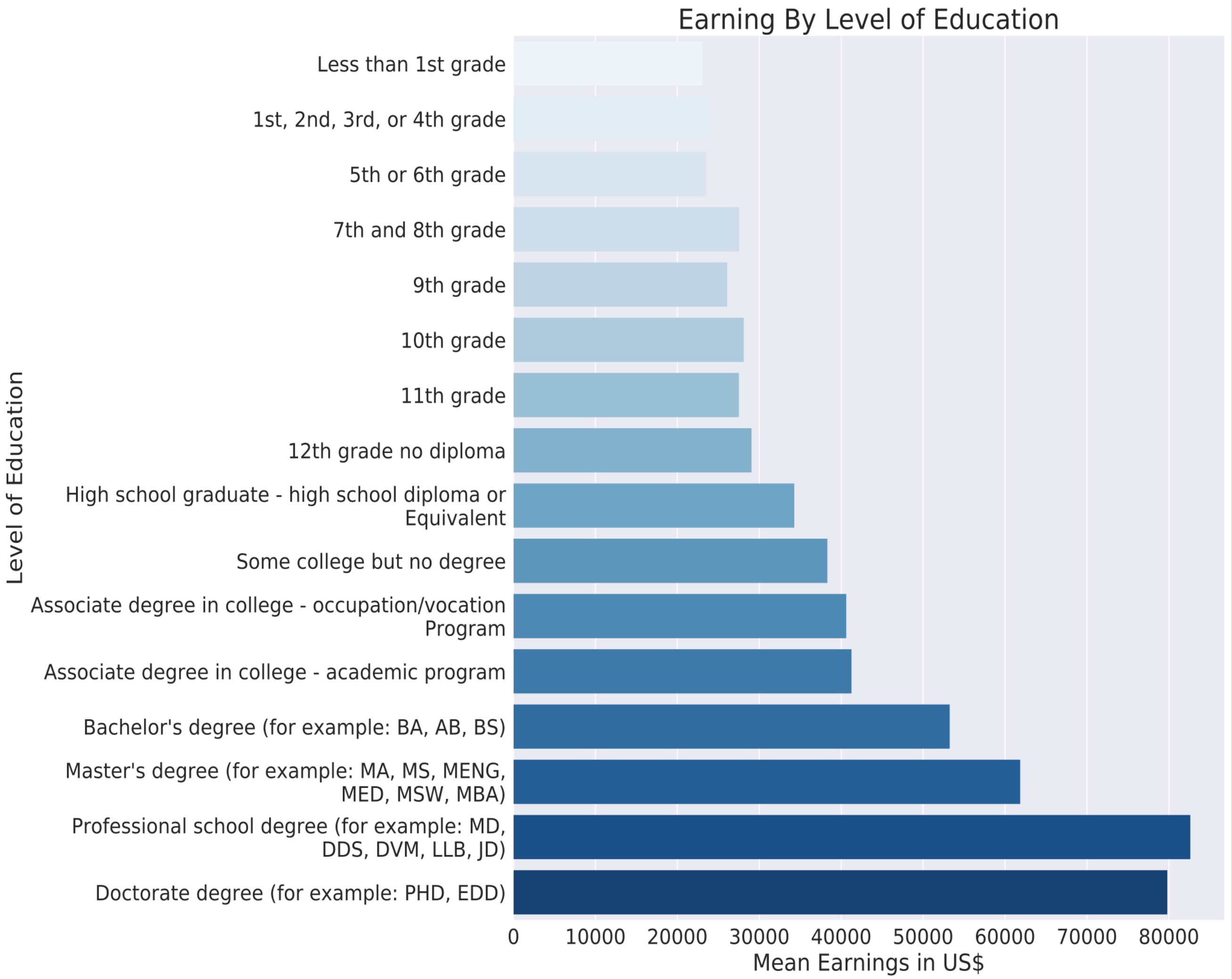
- *Statistically significant difference in the mean between states.*



2007 USA Mean Income By State
Population Mean = $41.95k
STD = 3.92k

What affects our salary?

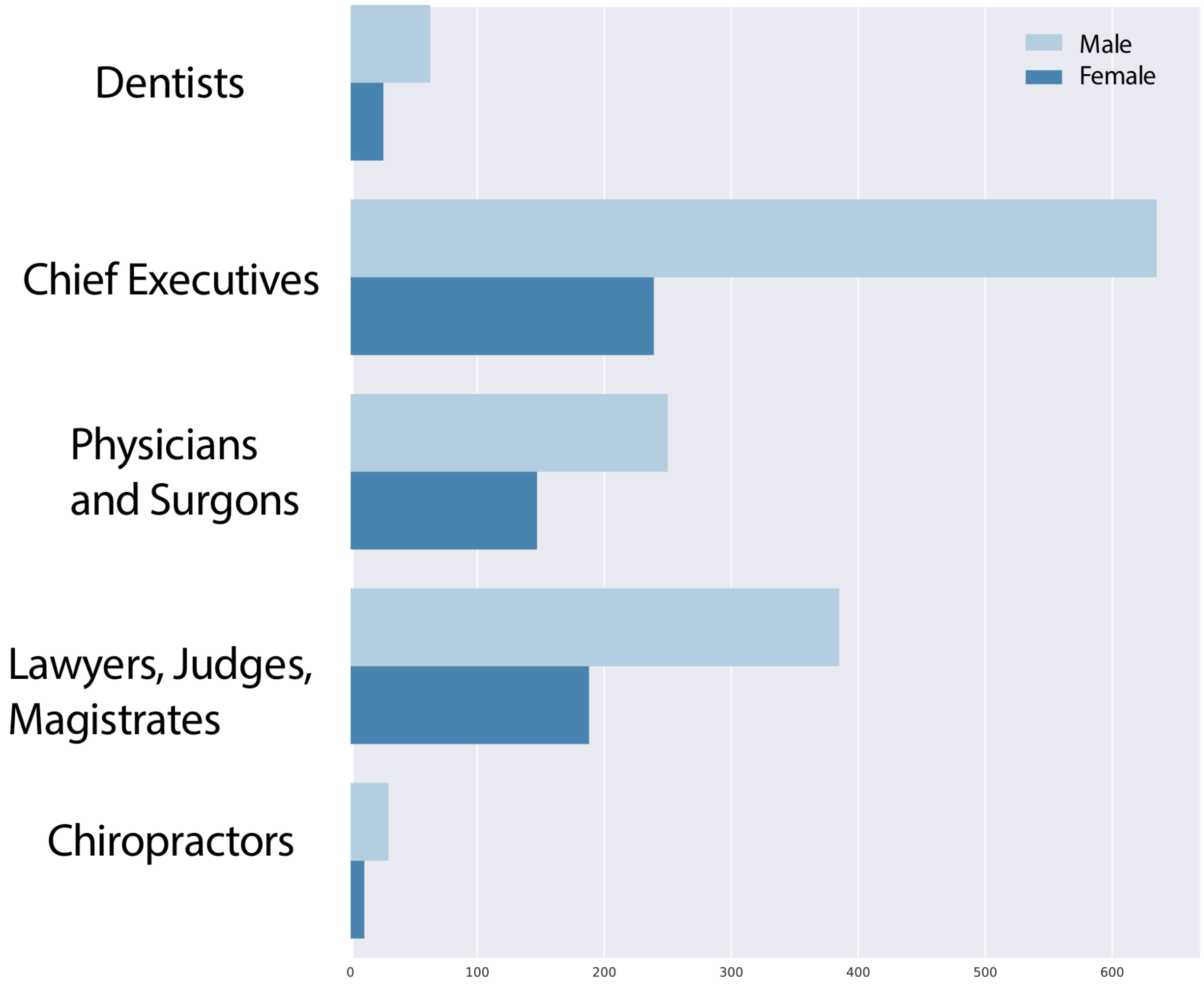
Education



- Education level is correlated with mean income.
- *There are distinct jumps for high-school graduates, college graduates and post-graduates*
- *Statistically significant difference in the mean college graduates and non graduates.*

What affects our salary?

Gender



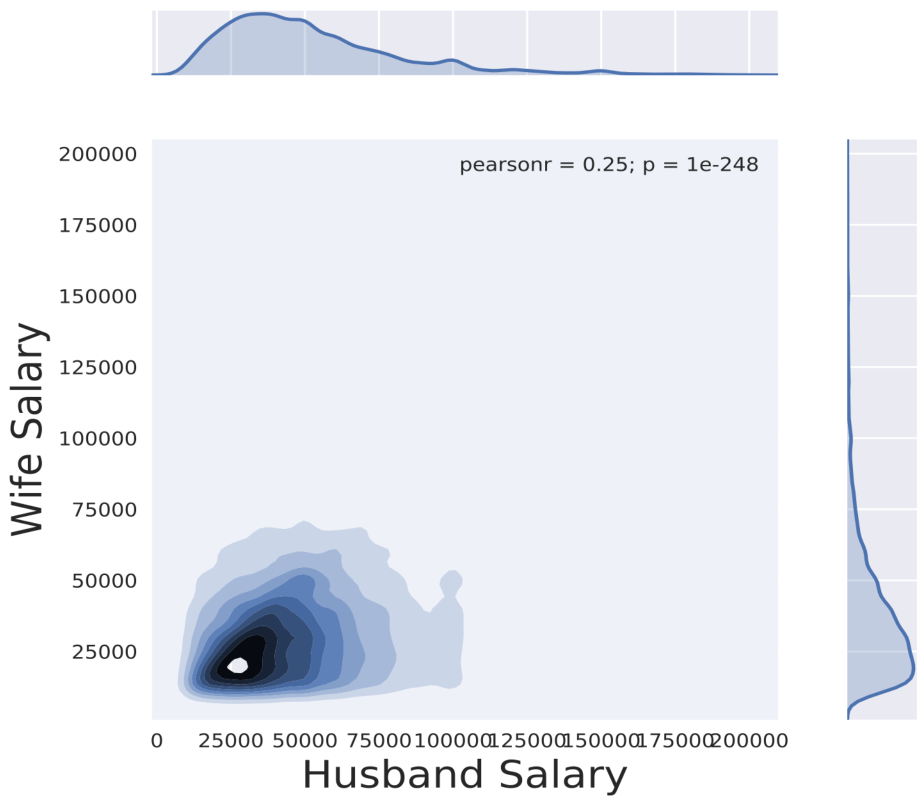
- Women are under represented in high paying occupations

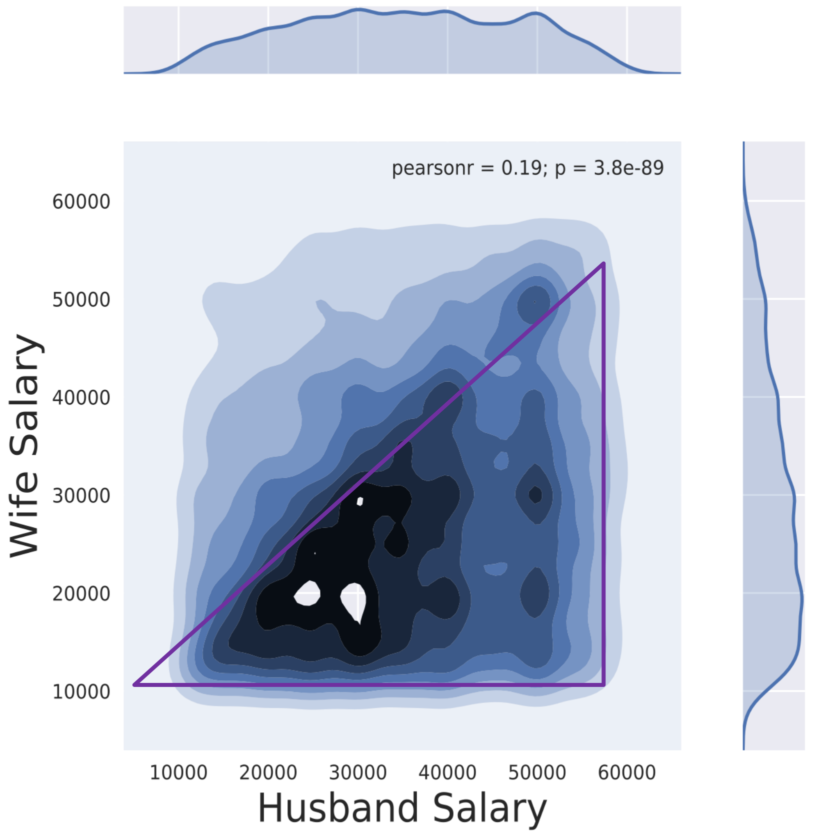
- *There is also a difference in pay scale within the same occupation (not general for all occupations)*

- *Women tends to earn less then their spouses within the same household*

What affects our salary?

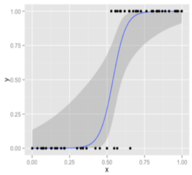
Household correlations

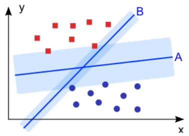


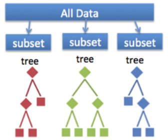


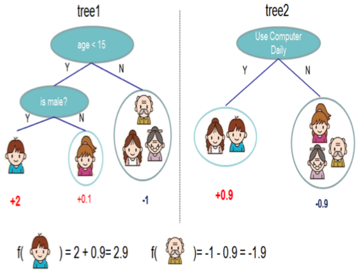
Classification

Comparing classifiers







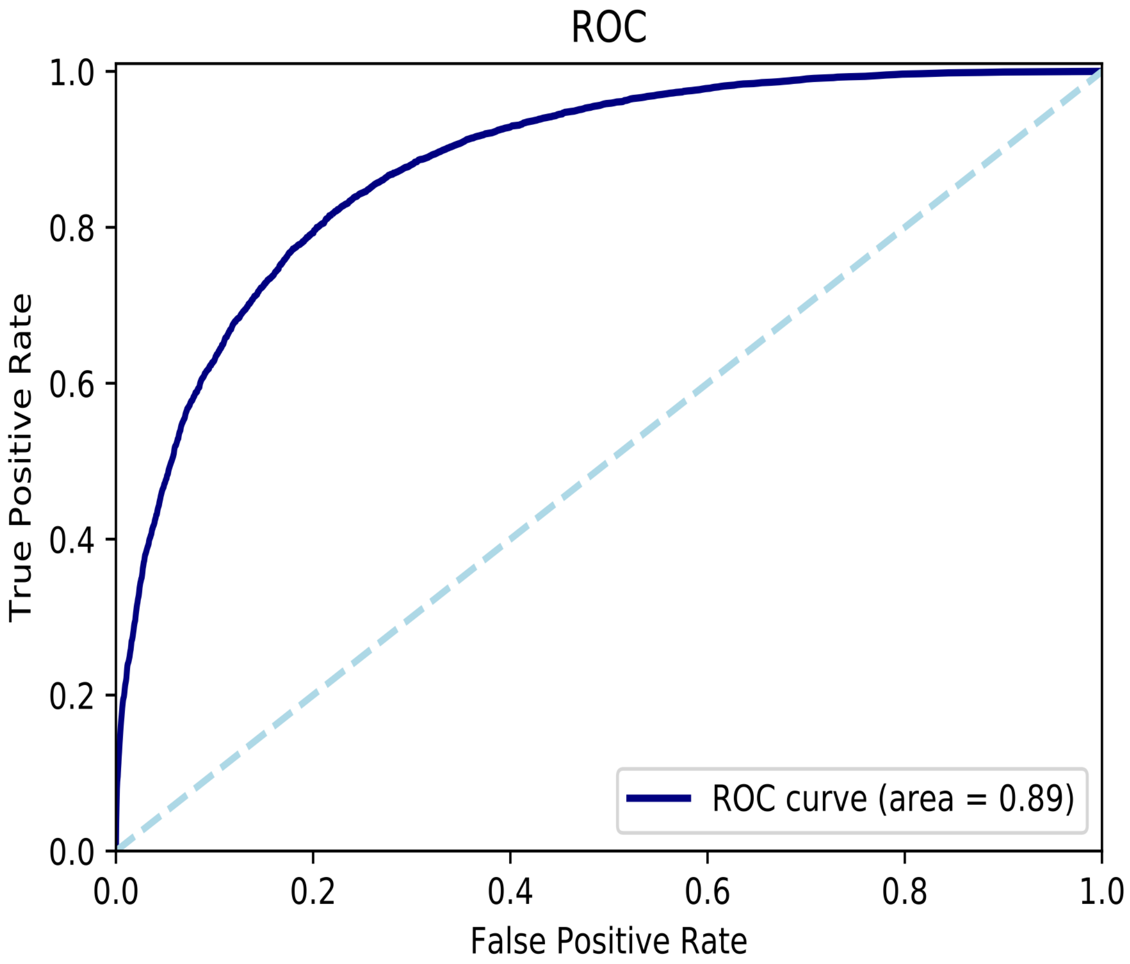


Classifier	Accuracy	Cohen's Kappa	F1-Score
Logistic Regression	0.78	0.55	0.722
SVM	0.8	0.55	0.719
Random Forest	0.81	0.56	0.71
Gradient-Boosting	0.81	0.55	0.719
XGBoost	0.81	0.58	0.724
lightGBM	0.81	0.56	0.74

- Logistic regression is fast and give good f-1 score
- *SVM is has higher accuracy but worse f-1 score ("accuracy paradox")*
- *For random forest we get better kappa score, but worse f-1 score, better true negative detection*
- *Gradient boosting performs really well. XGBoost is best but LightGBM gives similar prediction levels and is much faster*

Classification

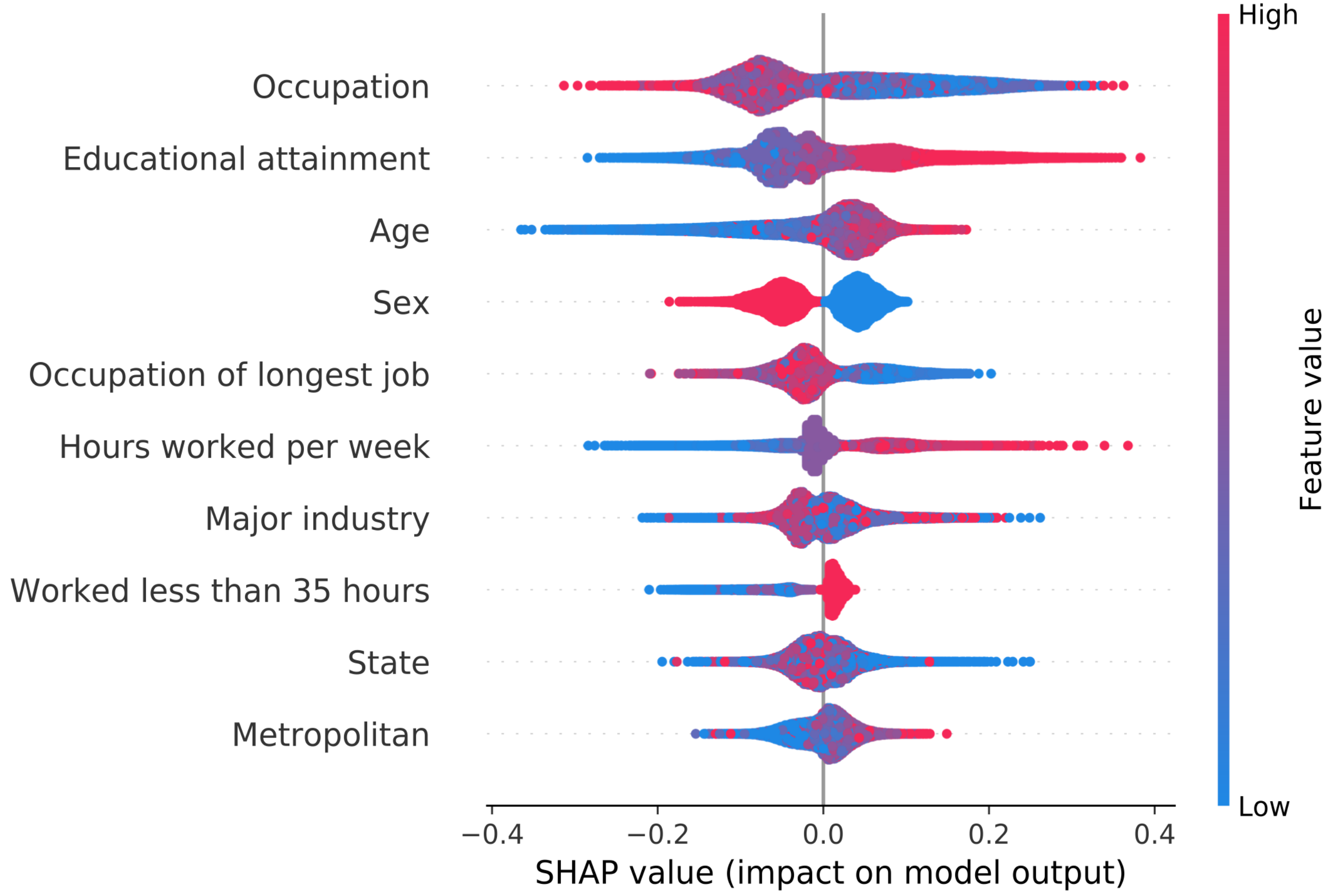
ROC curve and AUC



- ROC curve for LightGBM
- *The AUC is 0.89*
- *The threshold should be tuned according to the importance of the true positive rate*

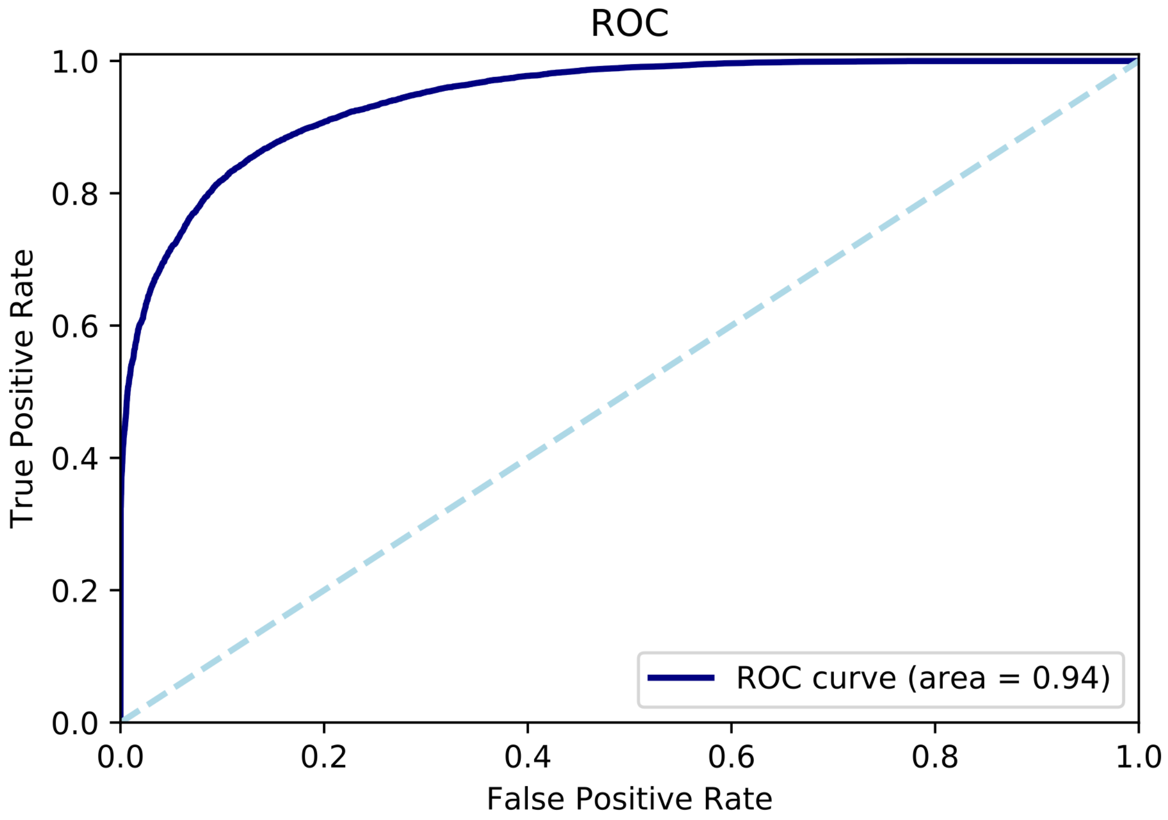
Classification

Feature importance



Classification

Predicting gender



- The classifier can be used to accurately predict gender

- *f-1 : 0.85*
- *Cohen's kappa : 0.72*
- *Accuracy : 0.86*

- *Discriminative job market*

Suggestions for future improvements

- Some of the occupations are underrepresented. Getting a more accurate distribution of salary for each occupation should improve the score.
- Grouping together similar occupations (or occupations with similar wage distribution)
- Sample microdata for the US Census are readily available online and contains millions of records
- Scarping Glassdor™ for income distributions
- Multi-Class classification methods