Predicting Lower Income Clients Based on the 2007 United States Department of Commerce Current Population Survey

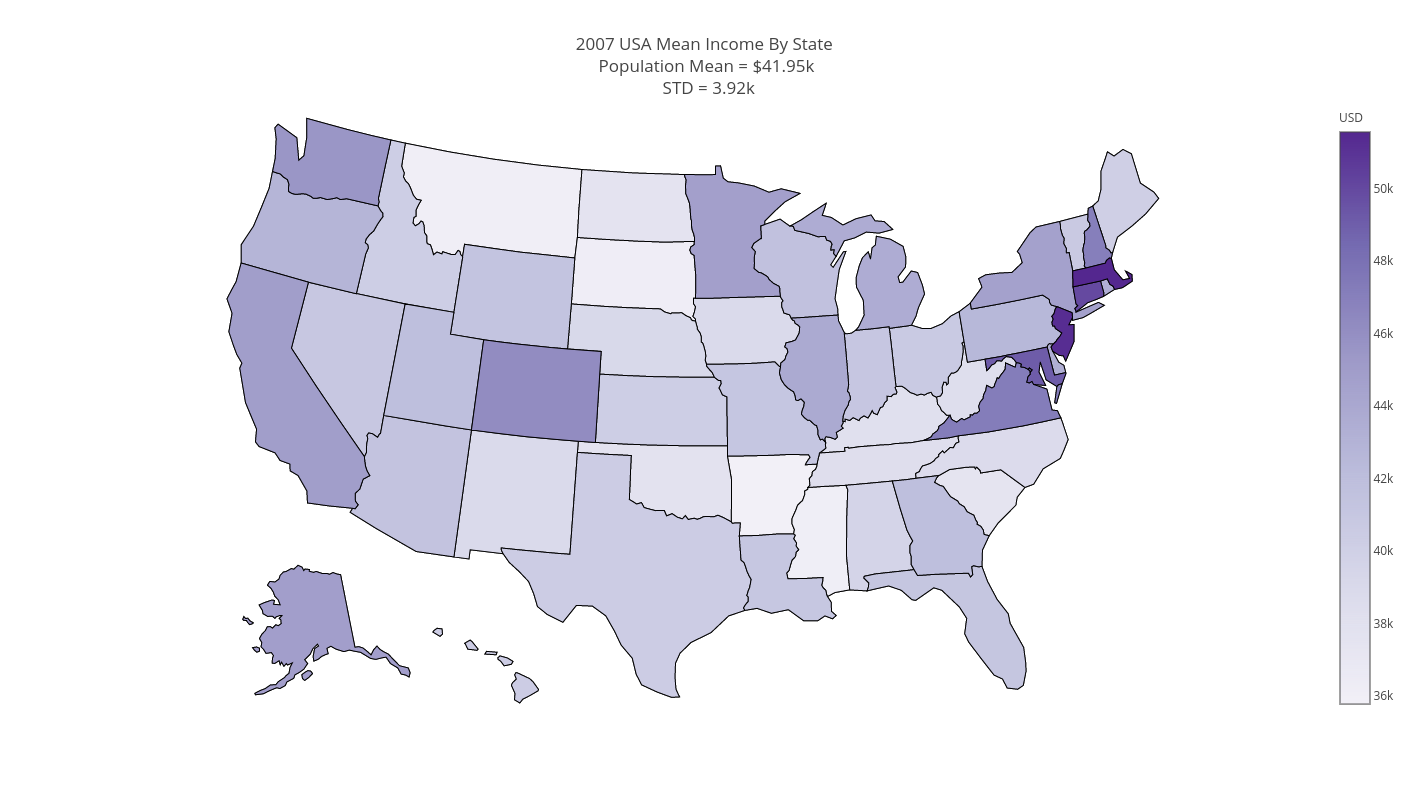


Table of Contents

[Introduction 3](#_Toc513155081)

[Definition of the problem 3](#_Toc513155082)

[Possible clients: 3](#_Toc513155083)

[The Dataset 3](#_Toc513155084)

[Obtaining the data 3](#_Toc513155085)

[Data wrangling 3](#_Toc513155086)

[Cleaning the data 3](#_Toc513155087)

[Examining wage distribution by occupation 4](#_Toc513155088)

[Other possible data sources 4](#_Toc513155089)

[Data analysis and inference: What affects our salaries 4](#_Toc513155090)

[Salary distribution 4](#_Toc513155091)

[Occupation 5](#_Toc513155092)

[State 5](#_Toc513155093)

[Level of education 7](#_Toc513155094)

[Sex 8](#_Toc513155095)

[Correlations in spouses’ salaries 9](#_Toc513155096)

[Predicting lower income individuals 11](#_Toc513155097)

[Linear regression 11](#_Toc513155098)

[Classing income level lower than $40K 11](#_Toc513155099)

[ROC curve and AUC 13](#_Toc513155100)

[Feature importance 14](#_Toc513155101)

[Predicting a person’s sex 15](#_Toc513155102)

[Suggestion for future improvements 16](#_Toc513155103)

# Introduction

## Definition of the problem

The project objective is to predict lower income individuals based on the *“2007 Department of Commerce Current Population survey”*. The cutoff salary for classification was chosen to be $40K per year. This salary is about the mean salary in the U.S (as of 2007) and about four times the poverty line.

## Possible clients

The data will be useful for credit companies in the risk assessment stage. While considering a long-term loan, being able to differentiate between lower and higher income and not just depend on current wages, which can be transient anomalies, can be very useful. Specific clients included companies like: Ernest (https://www.earnest.com/) and Upstart (https://www.upstart.com/)

# The Dataset

## Obtaining the data

The data for the project was obtained from the *“2007 Department of Commerce Current Population Survey”* (for which the raw non-aggregated dataset is readily available for academic research). The dataset consists of approximately 200,000 individual records obtained from 60,000 households, from which approximately 100,000 belongs to working adults. The dataset includes about 700 sociological markers (features), but only 46 have less than 30% missing values.

## Data wrangling

### Cleaning the data

To make the input file a bit smaller I stated by filtering out several of the obviously unnecessary columns (survey control number, month of survey, which is constant etc.) I had also filtered out records belonging to children under 17, and people over 80. Both populations are unlikely to have a steady job and are not in the interest range of the client. I have also filtered people who are currently unemployed. After consideration I also decided to exclude people earning more than 250,000 per year, as they are scarce in the data. While being explained outliers (their share of the sample matched the general population) there are just not enough samples to reach any conclusion about this particular group which seems to have statistics of their own.

For easier graphical analysis that states fips code from the data had to be converted to state names. This was done with a dictionary obtained from the census bureau website.

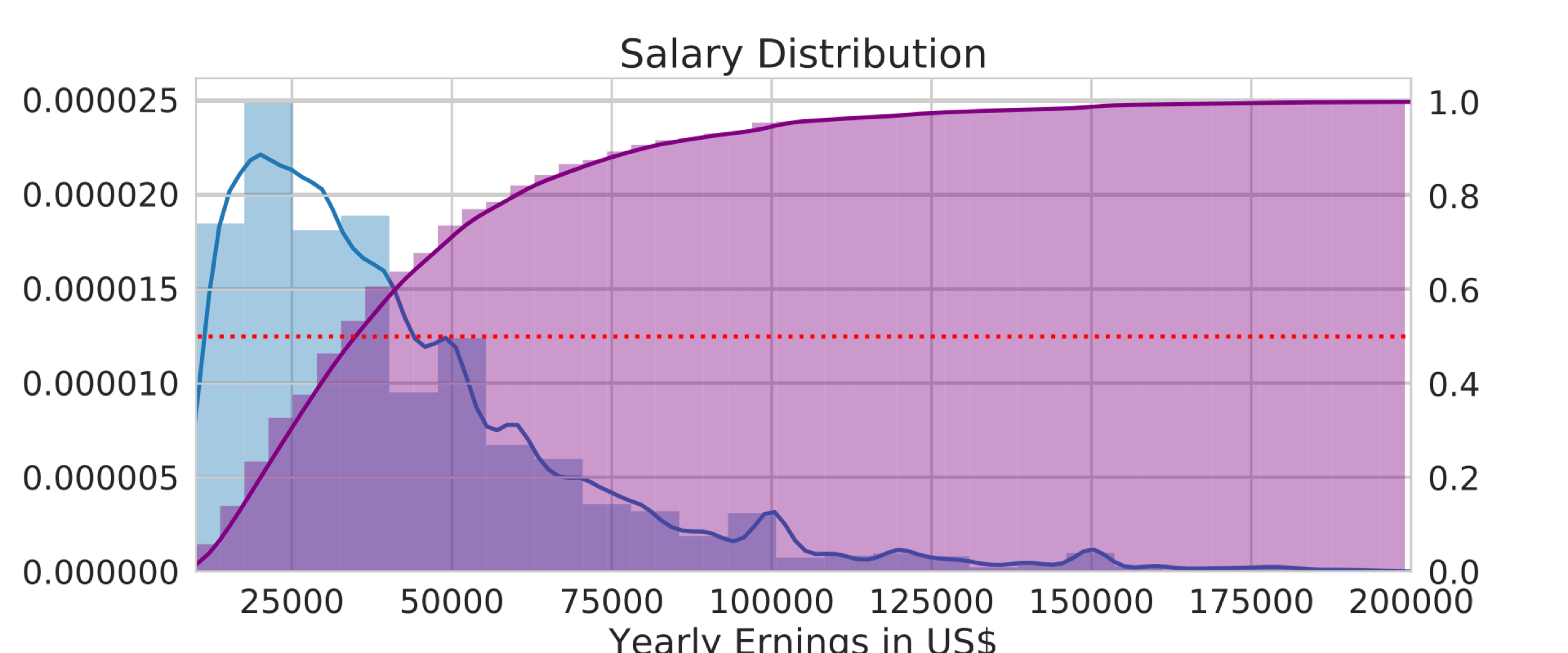
### Examining wage distribution by occupation

Examining the wage distribution by occupation shows some outliers. while some of them are probably created by sampling the distribution tails and not having enough samples, some seems to be just wrong (i.e. elementary school teachers making six figure salaries). These cases are rare and needed to be filtered. For example, one of the higher earning people is a butcher who was making $600K per year, while the mean for that occupation is around $30K. This may be a typographical error. It was decided to filter individuals earning more than three standard deviations from their occupation mean were also filtered. The filtered data was around 1% of the data and the prediction improvement was significant. Other possible data sources

# Data analysis and inference: What affects our salaries

## Salary distribution

As salaries over $200,000 are sparse we keep only those individuals who earn between $10,000 and $200,000 per year. Below is a histogram and a CDF of the salary distribution over the filtered population:



There are 80859 individuals who meet the above criteria. We can see that the mean salary in this group is: $42,703 per year, while the median is much lower: $35,000 (as expected observing the histogram above).

## Occupation

Obviously, a person occupation will have a significant impact on the salary he earns.

Below is a plot with the five most profitable occupations. As expected the mean income for those occupations is much higher than the general population, and we can safely assume that physicians for example will earn more than $40K per year. As the mean salary for many other occupations is around $40K per year with high standard deviation, we cannot make the classification based on occupation alone.

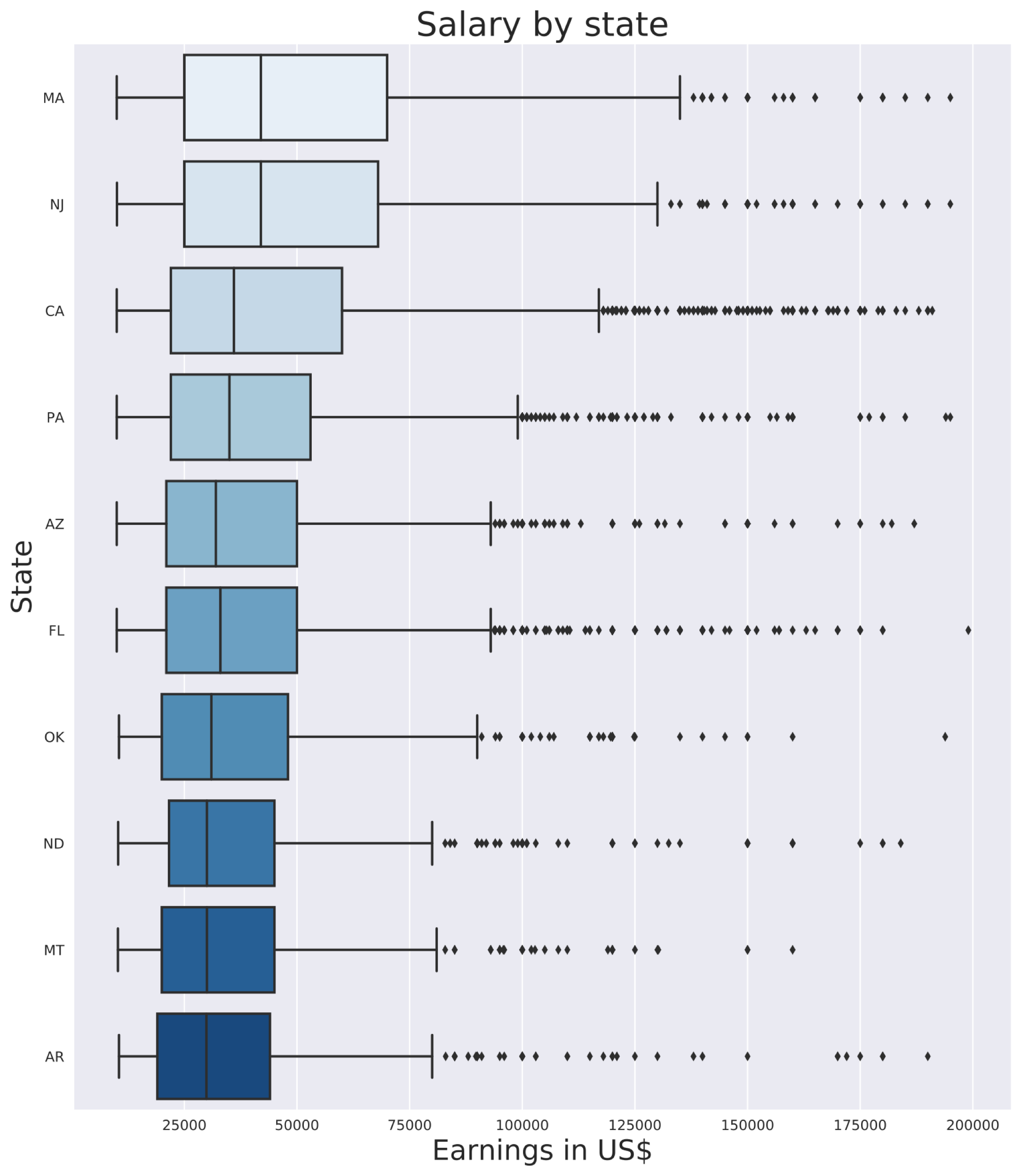
Below is a graph of the five occupations with the highest mean salary according to the dataset:



## 

## State

We can now explore the salary distribution by state. The boxplot below groups salary by state (for brevity it includes only ten states, the full plot is included in the milestone report) The states are ordered such that the state with the highest mean salary (MA) is at the top of the plot and the state with the lowest mean salary (AR) is at the bottom.



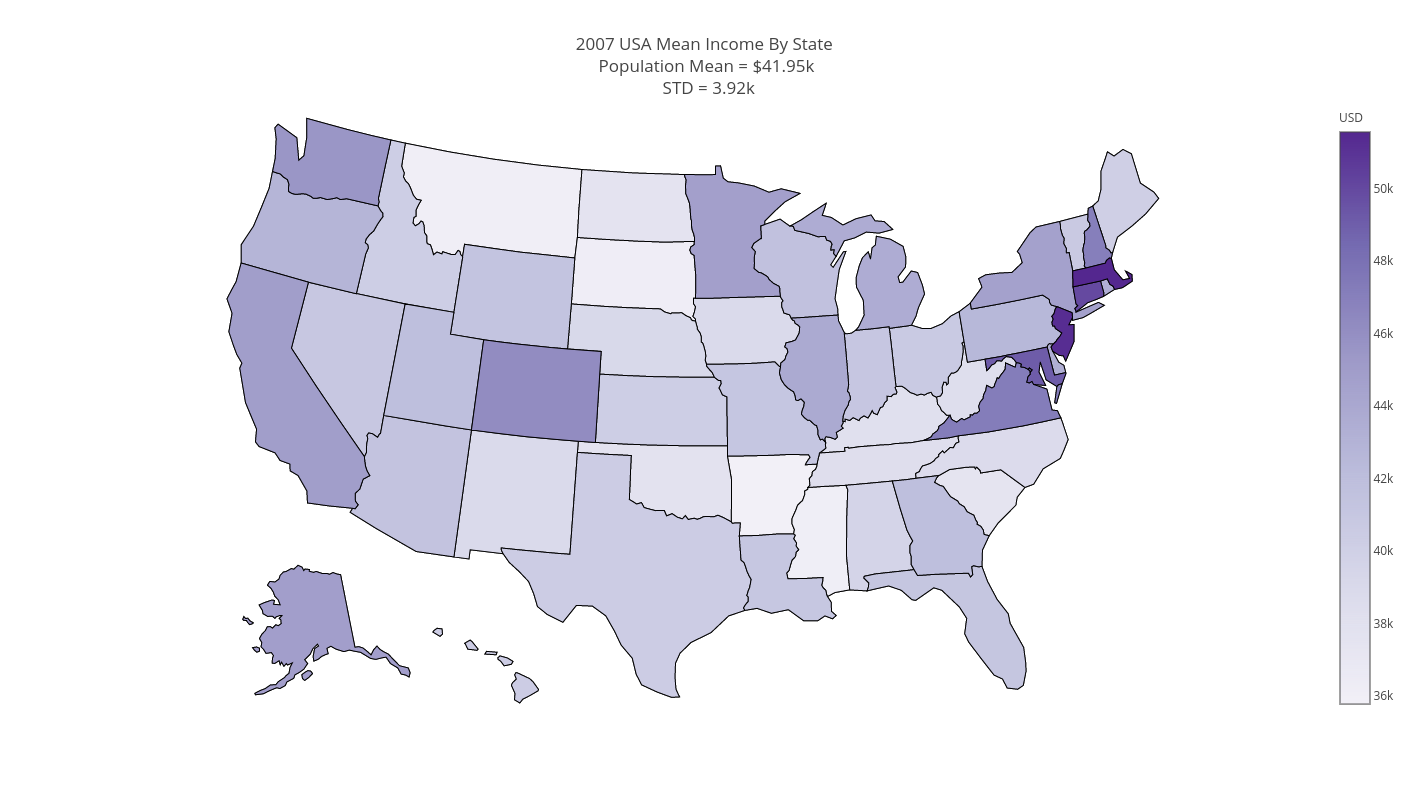
There seems to be a significant difference between states. Some other insights we can get from the plot are for example:

1. Salaries above $100K are not even outliers in MA and are very sparse in AR .

2. While CA is not at the top of the list (but not by far) in terms of mean salary, it does seem to have a very dense outliers earning significantly higher than the population median.

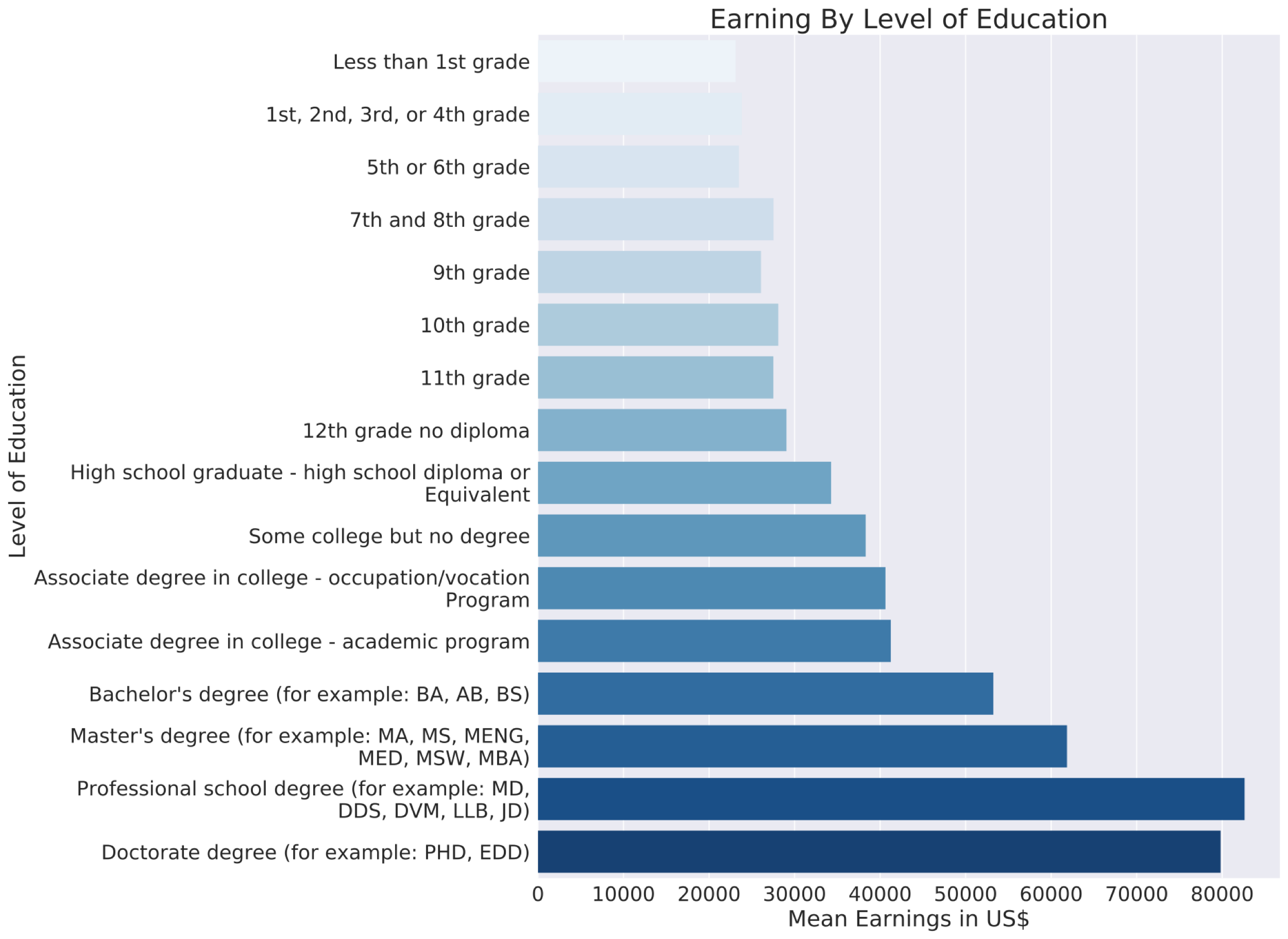
3. There are no outliers at the lower end of the salary range (this is apparent from the salary distribution in the population, and also happening if we do not cap the lowest salary at $10K per year)

The difference between the mean income in MA and AR is $15789.774. To check if this difference is statistically significant a *t-test* was performed. The *p-value* obtained from the test was: 5.2x10-29, this is an extremely low value which infers that there is a difference in the mean salaries between the states.

The heatmap below shows the mean salary of each state in the U.S:

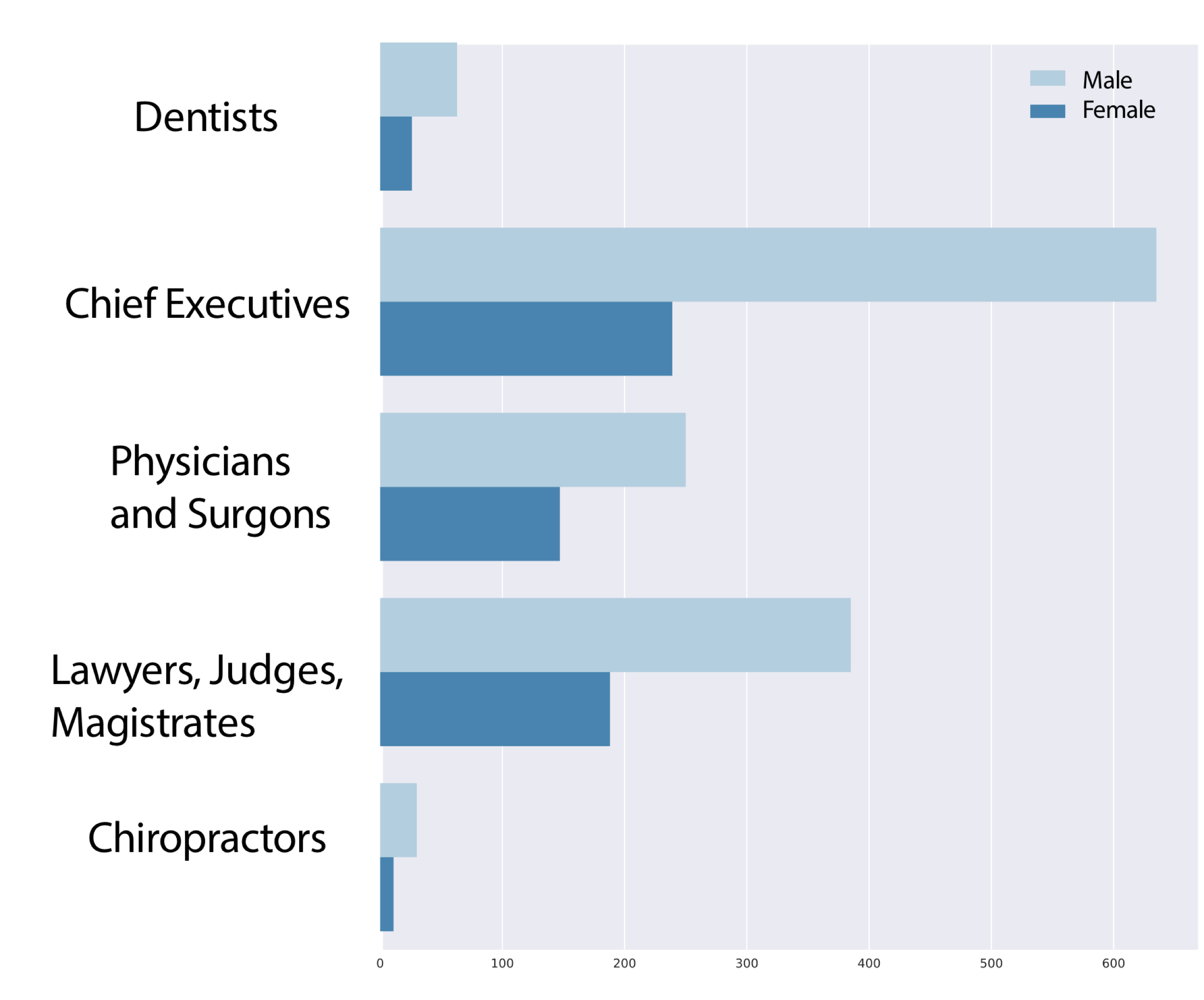
## Level of education

We can expect that on average, higher level of education will lead to higher salary. The difference in the mean salary of college graduates and non-graduates is: $19538.938. The *p-value* for the difference in means is practically zero, with a *t* statistic of 1.01x102. This gives us extremely high confidence in this result. The figure below shows the mean income by education level. It clearly shows that the higher the education level, the higher the mean salary. There are distinct jumps for high school graduates, college graduates and post graduates.



## Sex

Unfortunately, we found that there is a significant difference between the mean salary of females and males. The mean difference is $12172 per year, with a *p-value* which is practically zero (the *t* statistic is 63). Part of this difference is due to the fact that women are underrepresented in high earning occupations. The following plot shows the ratio of men and women in the top five paying occupations:



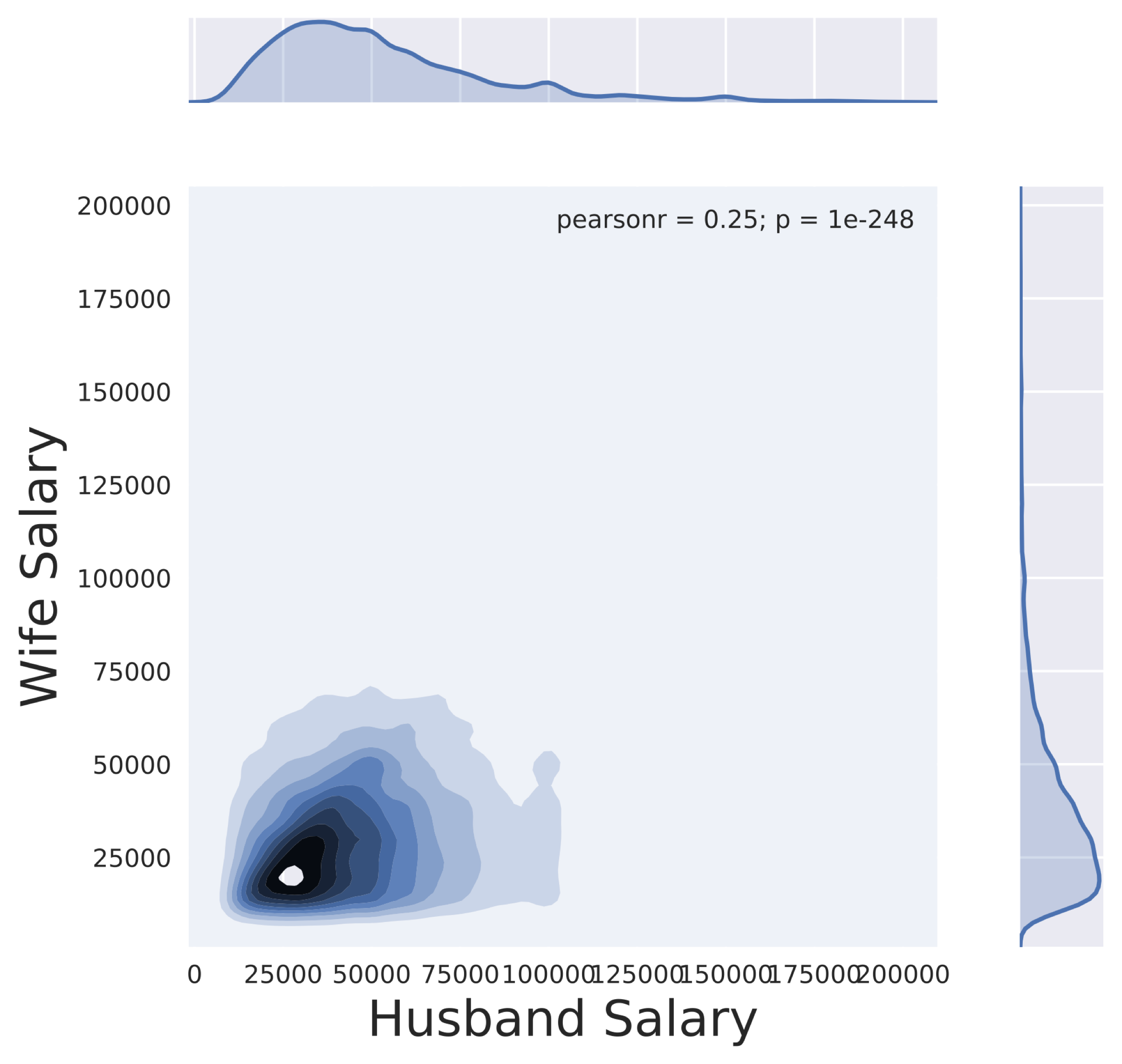
The imbalance is strikingly apparent, especially among Chief executives.

While this imbalance can explain some of the salary differences, we have found that even within the same occupation there is a difference in salary between men and women.

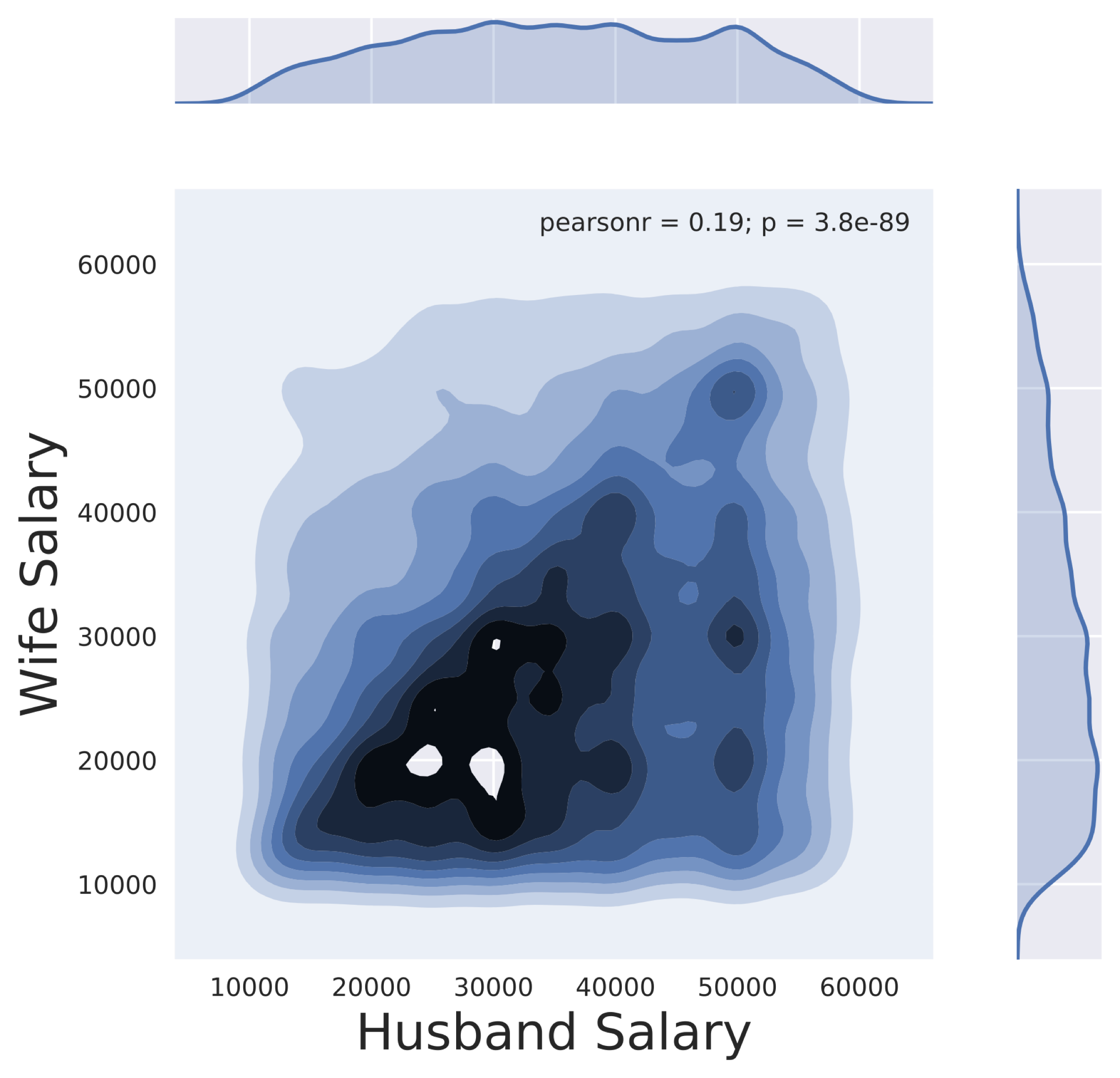
For CEO's, for example, we get that there is a significant difference: $21618.672 per year with a *p-value* of 5.6x10-11. In other occupation the situation is better, for example for computer programmers the difference in not statistically significant, with a *p-value* of 0.3

## Correlations in spouses’ salaries

In addition to individual records the dataset contains household information. This information was used to test the hypothesis that spouses’ salaries are positively correlated. The dataset was filtered again to contain only households where there are two working spouses. The following joint plot shows a positive correlation between the salaries of spouses, with a Pearson correlation factor of: 0.25.



A joint plot which is limited to salaries between $10K and $60K, which contains most of the mass of the distribution is shown below:



The correlation is more apparent in this plot (even though the Pearson correlation coefficient is smaller). The heavily populated lower triangle shows that even within the same household men are earning more and are the main providers.

# Predicting lower income individuals

## Classing income level lower than $40K

We continue by converting the problem into classification problem. A salary cap was set at about four times the poverty level ($40K per year) and a True class was assigned to each person earning less than this threshold. The resulting dataset is slightly unbalanced, with 64% of the samples in the True class. The features used for classification were chosen such that they can contain information which can be readily collected by the client, and their value changes over the years can be easily speculated, so the client can make informed decision based on projected income level. features like age, education level, sex and state were included but capital gain and loss which are features with obscure feature value which is hard to predict. A grid search was then used to find optimal parameters for 6 different classifiers. The grid search was performed on a computer cluster, using five different nodes with 32 Intel Xeon Phi processors allocation on each of them. The following classifiers were fitted:

1. Logistic Regression classifier
2. Random forest classifier
3. SVM classifier
4. XGBoost classifier
5. Gradient boosting classifier
6. LightGBM classifier

A pipeline was produced to OneHot encode the categorical features and to scale the numerical features (expect from LightGBM which does it very efficiently by itself).

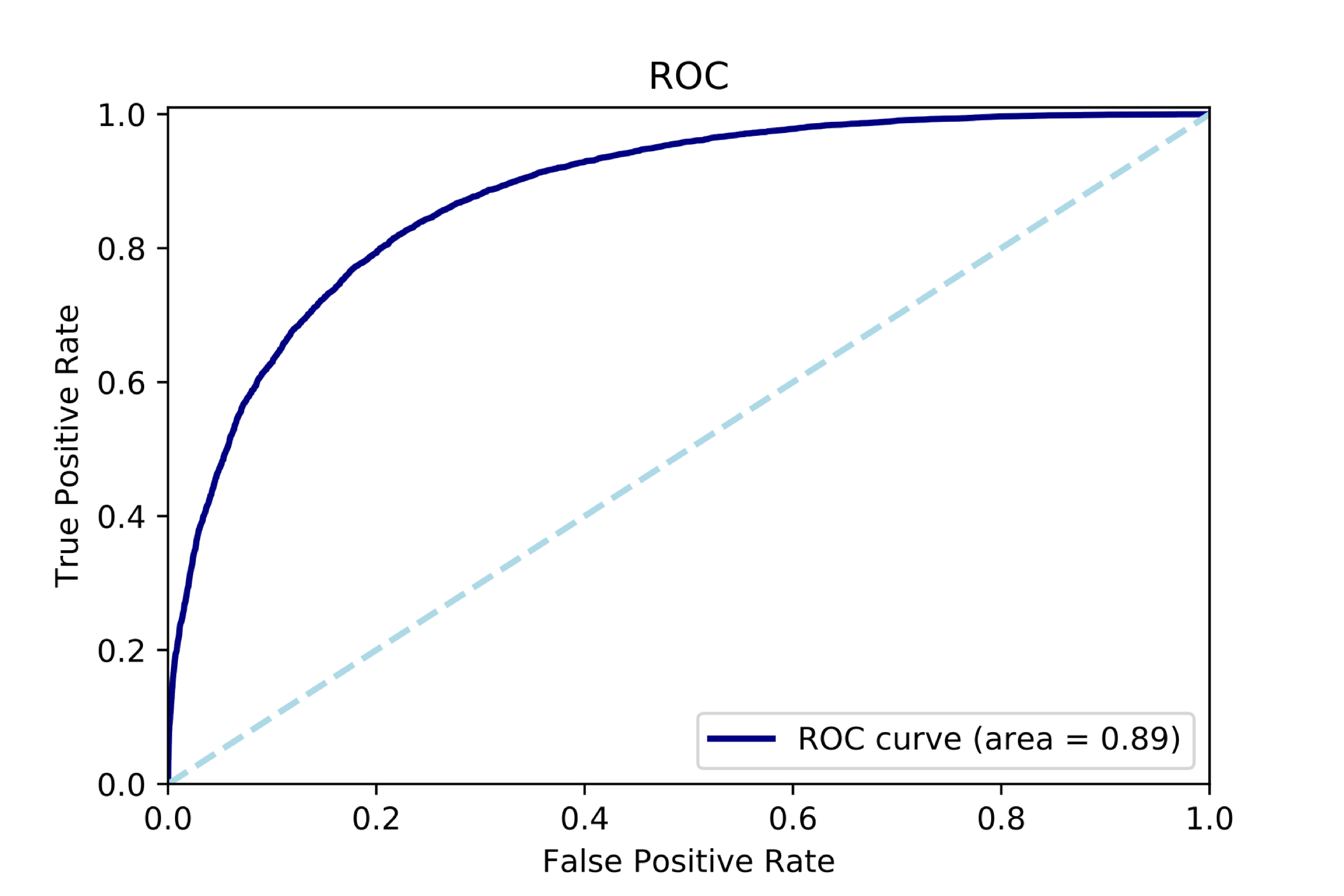
Logistic regression was chosen for its fast performance and was used as a reference point for the rest of the classifiers. SVM is expected to handle well higher dimensional problems (which is the case for our classification problem due to the OneHot encoding), but scales poorly with the dataset size. Random forest is a relatively fast algorithm and is less prone to overfitting then just a single decision tree. The last three are gradient boosting classifiers which are current state of the art and are expected to preform best. While XGBoost usually perform better then LightGBM, LightGBM runs faster (as it is bucketing continuous features) and allows to grid search over larger parameter space.

A fivefold cross validation was used during the grid search and the following table reports the classier score on a test set which contains 20% of the original data which was separated before the fitting procedure (test set). The optimal parameters for each classifier are shown in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Accuracy | Cohen Kappa | F1-Score |
| Logistic Regression | 0.78 | 0.55 | 0.722 |
| Random Forest | 0.81 | 0.56 | 0.71 |
| SVM | 0.80 | 0.55 | 0.719 |
| Gradient-Boosting | 0.81 | 0.55 | 0.719 |
| XGBoost | 0.81 | 0.58 | 0.724 |
| lightGBM | 0.81 | 0.56 | 0.74 |

All classifiers perform significantly better than just consistently classifying all samples as True, which would yield an accuracy score of 0.65. While logistic regression performs the worst in terms of accuracy, both the kappa score and the f-1 score are relatively high, which is an example of the accuracy paradox which states that sometimes model with lower accuracy have better predictive power. Accuracy should be considered as a metric only if the weight of the true positives and true negatives is the same. For our purpose it is more important to detect the True class, which makes the f-1 score a better metric. The f-1 score does not take into account the true negatives and for this specific application, when it is more important to correctly classify low earning individuals it makes more sense to use it for comparison. If the true negatives are also important we should compare the kappa scores. The logistic regression classifier gives reasonable f-1 results and was trained quickly. If resources are limited this should probably be the classifier to use. SVM as expected scaled poorly and took large amounts of time to converge with no apparent improvements in performance. While the random forest performs well when parallelized also fitted the data quite fast, the gradient boosting classifiers outperformed it. From the three gradient boosting classifiers, XGBoost and LightGBM are considered state of the art. XGBoost achieved the highest kappa score, which will make it the one to use if we care about the true negatives. LightGBM obtained the best f-1 score presumably as a result of the larger parameters space explored. As the training speed is excellent and the f-1 accuracy achieved the best of all classifiers tested this is the classifier which is recommended to use (the fitted classifiers along with the best meta parameters are in my GitHub account in the corresponding folder).

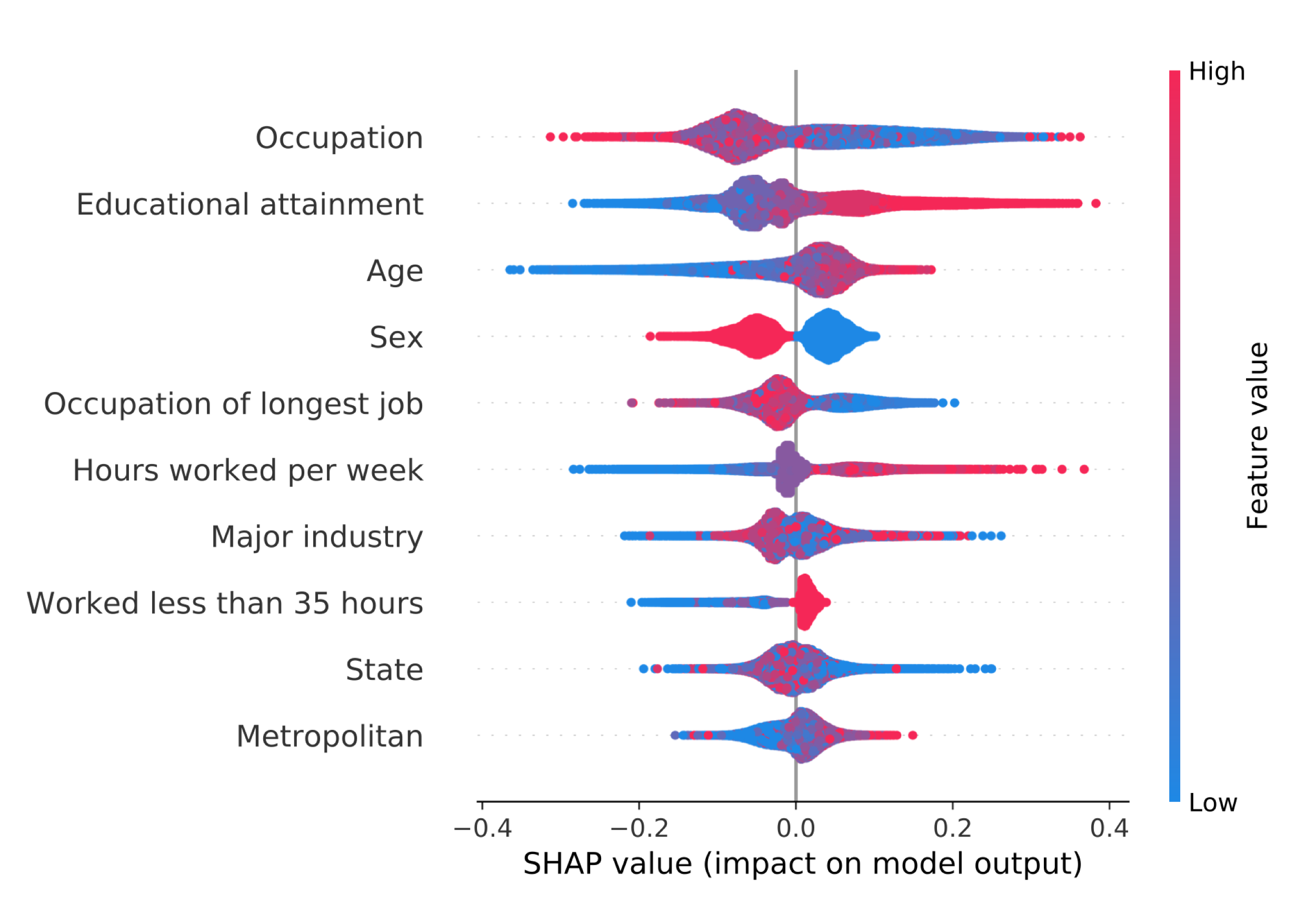
## ROC curve and AUC



The AUC score for the chosen classifier is 0.89 and is a significant improvement from the score a random classifier will achieve (0.5). Depending in the use of the by the client (presumably as part of a larger pipeline) the rate of false positives should be adjusted to an acceptable level (by changing the classification threshold). For example, of a false positive rate of 0.2 we will get more than 0.8 true positive rate. If the output will be further used in a pipeline, it will probably make sense to output the classifier probabilities for each class rather than the classes themselves.

## Feature importance

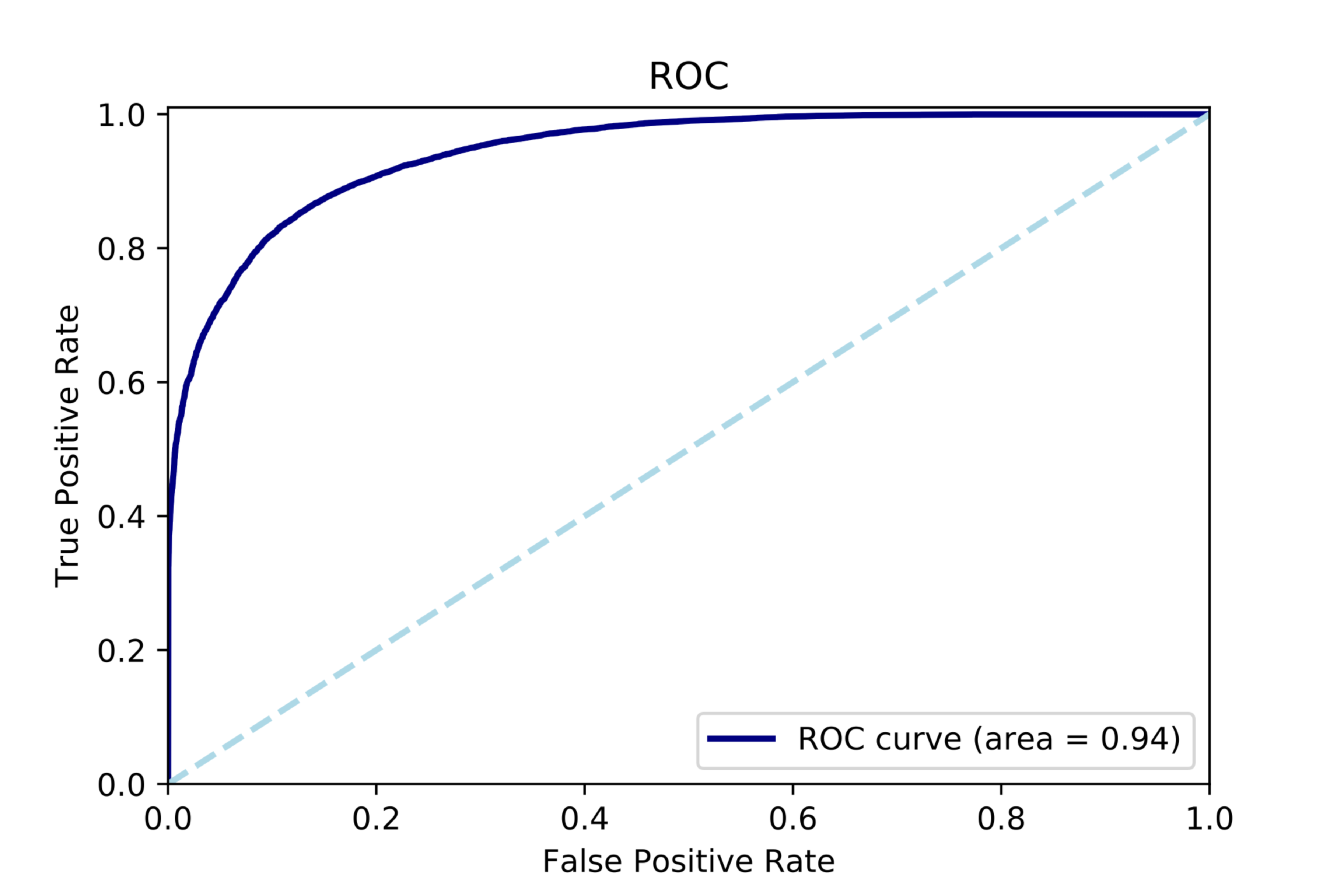
Another useful information we can obtain from the fitting procedure is feature importance. It can give an insight as to which attributes will affect a person’s salary and could help decide which information is needed from the applicants to get the best estimation of their potential earnings. While LightGBM provides a method of obtaining feature importance, it was found that stochastic tree methods in general will not give a consistent score for each feature due to the inherent randomness and greediness of the algorithm. A recent paper (<https://arxiv.org/abs/1802.03888>) suggest a method named Tree SHAP (SHapley Additive exPlanation) which can provide a consistent and “demonstrate better agreement with human intuition through a user study”. The authors integrated their library into XGBoost and LightGBM and will be used here to explore feature importance. Below are the ten most important features obtained from the Tree SHAP method (applied to the LightGBM classifier):



The plot contains an abundance amount of information. As expected the occupation has the highest impact on the salary earned. We have also seen in the previous section that educational attainment and sex have a large impact on a person’s salary. We also notice that the age has a significant impact on the salary, mainly a negative effect for younger individuals. While features that are categorical have random high and low value in the plot as the distance between them have no meaning, ordinal features such as age and level of education are even more interpretable. We can see for example that lower education have a higher negative impact on the score, while higher values of the education number are exclusively related to increased score. It is also shown that the sex has negative effect for higher values (which mean females as they are marked as 1 in the data), which was obvious in the statistical analysis.

## Predicting a person’s sex

The high imbalance in earnings and occupation representation between male and female suggest that the classifier can do a good job predicting the sex of an individual. To test this hypothesis, we have applied the same classifier to the dataset, but switched the classes to be male/female and the salary was introduced as a future. Below is the ROC curve obtained from the fit:



With an AUC of 0.94 and f-1 score of 0.84 the classifier is doing a great job in predicting the sex of a person. I assume that further optimization of the meta parameter to the new problem will give even better results. While this is not one of the objective of this project in my opinion the above result is meaningful and should not be taken lightly.

## 

## Suggestion for future improvements

While the classifiers did a decent job in predicting lower income they were far from perfect. Following are some suggestions for future work and improvements:

1. Some of the occupations are underrepresented. Getting a more accurate distribution of salary for each occupation should improve the score. With the dataset we have now, perhaps grouping together similar occupations (or occupations with similar wage distribution, perhaps using a clustering method) could improve accuracy. Identifying occupations with similar salary distribution can also be done by scarping sites like Glassdor if it is hard to deduce the distribution from the data.
2. Obtaining a larger dataset will probably also fix the above problem. Sample microdata for the US Census are readily available online and contains millions of records. As most of the analysis was done on my personal computer it was proven to inefficient to work with the large dataset. Down sampling it would have introduced a sampling bias. In the smaller dataset I worked with this bias was minimized by survey methodology experts.
3. Even though the parameters for each classifier were optimized, it was by all means not an exhaustive search. A randomized search on a larger parameters space may reveal better meta-parameters.
4. By using multiclass classification methods (which are implemented in the advances tree classifiers and can be easily introduced to logistic regression and SVM by one vs. all or one vs. one methods) we can divide the data to more classes and classify for example low-income, mi-level and high-income individuals. With a more accurate classifier we may be able to predict income percentile with acceptable accuracy. Outputting the probability to be in each percentile can be proven to be useful in further steps of the decision pipeline.