Individual Contribution Report

Income Analysis & Prediction

Tiriac Ioana-Raluca
(Team 44)

Reflection

Participating in the project “Income Analysis & Prediction” I was at first involved with the extended data exploration and analysis of the adult US Census data and finding appropriate visualizations to help us establish the relationship between the income of adults and other features of the data set.

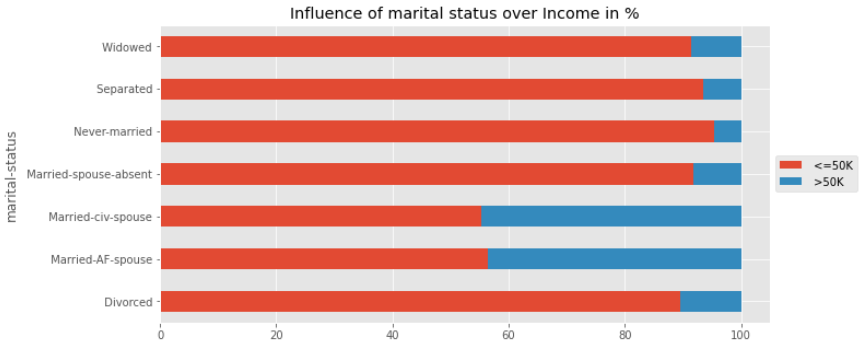
Later on in the next part of the project I learned about different possible classifiers used in machine learning and looked up a lot of material about how to implement the different algorithms – decision tree, gradient boosting and building an artificial neural network and decided to try out the gradient boosting classifier on the provided data set exploring the pros and cons of using it and later comparing it to other’s colleagues approaches like the decision tree classifier so we’ll be able to decide what has a better precision accuracy and how we’ll get a better improved feature correlation.

I also took a considerable amount of time to tweak our drafts for the presentation and documentation, bearing in mind what material we should present to the customer in the presentation without too much technicality but trying to remain professional and also focusing on the right structure for the documentation for it to display our work as engineers. I also invested some time thinking about “our brand” as a team – the “Data Vaders” and trying to convey that through our presentation video (hint: it’s the link in the presentation).

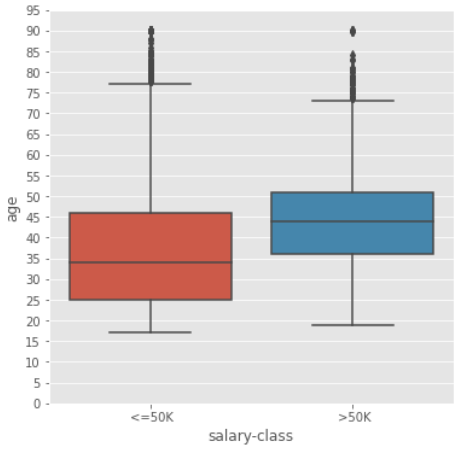
Lessons Learned

About visualizations

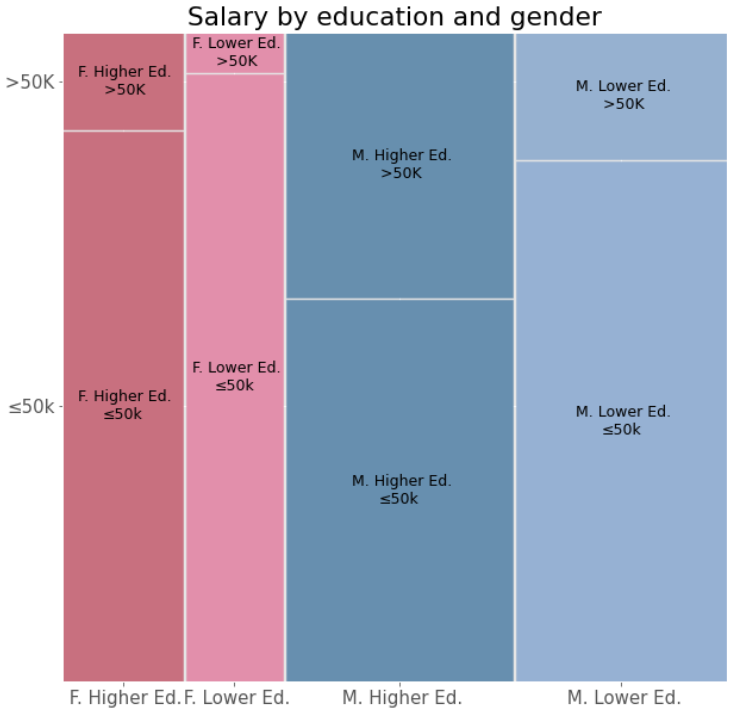
The first visualizations I did were horizontal stacked bar charts of all the different features in relation to the income, I found them to be more expressive, easier to read and compare than the vertical ones, but we later decided to use the latter because they took less slide space in the presentation than the horizontal ones (below an example)



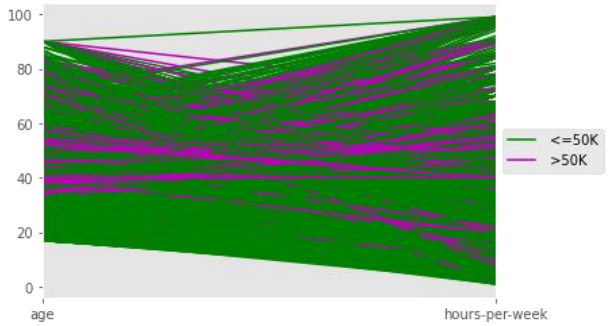
I realized how useful box-whisker plots are for comparing two data distributions and how much insight they can provide (done for salary-class and capital-gain), like comparing the average, min, max and outliers for both salary classes.

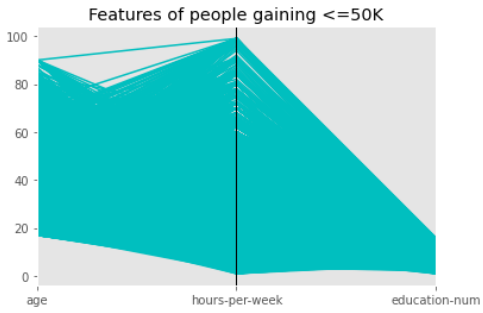


Later on in the data exploration I tried to focus on correlating 2 or more features with the income(salary) to draw conclusions on what impacts the salary class most. That's when I learned about creating meaningful mosaic plots using 3 variables (like gender, education, income) and obtaining a useful visualization for the given data set using custom colors. Customizing some details for some plots (like colors, labels and so in) was not as straightforward as I might have expected, like customizing the colors for a mosaic plot ☺



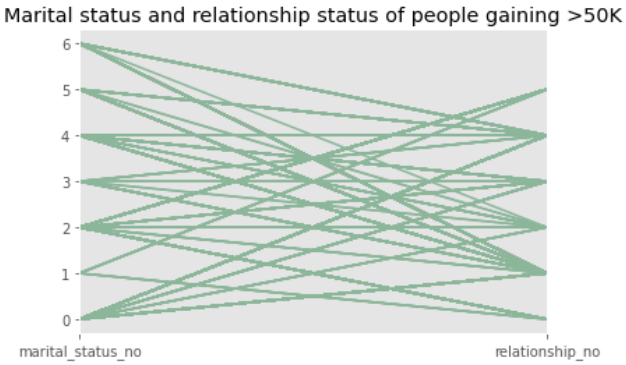
Among other visualizations, I explored the possibility of using the parallel-coordinate plot to compare more than 2 numerical features of the two different salary classes and then realized a couple of things like : how to turn categorical data into numerical one in Python/machine learning and the necessity of normalizing data. Below are some examples of this plot that the team later decided not to use in the presentation.





About data exploration

As mentioned before I learned about transformations data has to go through before a visualization is created, about data cleaning and normalization using Python. I also learned about turning categorical data in numerical data in machine learning, why it's important and the two possibilities: integer or label encoding and one hot encoding. I explored the label encoding when trying to use categorical data like marital status and relationship status on a parallel coordinate plot (the visualization did not turn out to be useful though)



Later I put the one hot encoding knowledge to good use by reviewing my colleagues' work about the artificial neural network he had trained.

About machine learning

Learned a lot about useful machine learning techniques, about hot encodings and the use of classifiers beyond prediction purposes, like running the model multiple times to achieve an average feature importance and not focusing so much on achieving a greater accuracy than 80%. Implemented and tried out the gradient boosting classifier on the given dataset, the model uses an ensemble of shallow trees and builds hundreds of them, therefore it was too challenging to visualize them so we (team) decided to use a decision tree that was easier to visualize if boundaries were used on it.

General

As general lessons I learned about synchronizing your work with those of other colleagues and why established deadlines are so important and have to be respected to come together as a team. I quite enjoyed our team's useful pair programming sessions and reviews of each other's work.

Assessment

I learned a lot from this team project in the area of producing visualizations using Python, but also a bit about how to start off with a data science project, how to break it down in smaller more manageable tasks and make engineering decisions along the way in cooperation with the team. I learned valuable lessons about what operations your data has to go through when preparing it for exploration (cleaning, normalization), how to start off with simple visualizations and decide which ones make sense to use, how to proceed with classifiers, models, how to train and test a model. I also got a chance to refresh my knowledge of training and testing a neural network learned in another class of “Artificial Intelligence” and deepened this knowledge by researching TensorFlow and its applications and know for sure I’ll be using this knowledge again since it’s in my area of study interest.

Future Application

I have drastically improved my Python programming skills and the way I look at big data sets, I have a pretty good idea of what packages and tools to use to produce different visualizations and what kind of visualizations to use for different purposes and to express certain data trends. This course has been a continuation for me in the vast area of big data and data mining but with a focus on visualization and I look forward to using this knowledge to enhance other projects I’ll be doing for work or university.