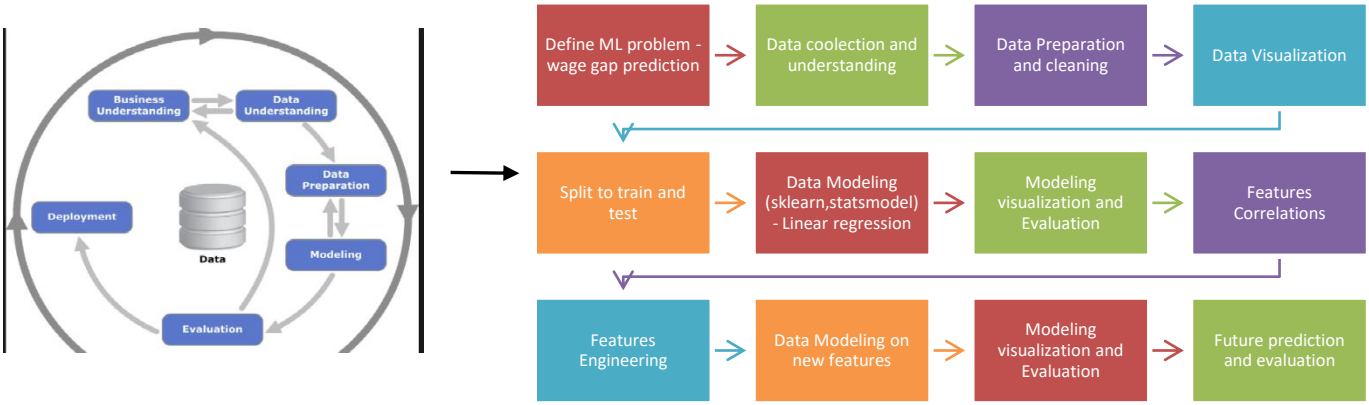
Data Science Workshop - Wage Gap Prediction

Introduction

With economic and cultural advances worldwide, we see that the wage and class differences between genders, are still substantial. In this project we will try to contribute to the cause of closing the wage gaps by finding irrelevant factors (features). We will do it by predicting the wage difference, focusing on the OECD countries. Our product will be a prediction for each state in the OECD for years to come. Our whole process, includes more than 1 iteration on the CRISP-DM steps. For some steps, we will elaborate on our different process stages. We use the model of (country, year) as a benchmark and test our results accordingally.

Jupyter notebook – We tagged each cell in our notebook with title and a number in order to refer you to the relevant cell. In order to make our notebook readable and comprehensive, and avoid code duplications, **we added help_module.py**. This module contain functions that create model evaluation graphs – prediction values vs data, average error and more. **In case you want to elaborate and see the implementation you can search in this module.**

Our work guideline was affected from the basic components of the data science process:



Business Understanding

In part of the world's progress in many fields, we expect the total gap average to diminish through time. In addition, we expect the gap to depend on parameters (features) which relate to population, GNP, and working population percentiles.
The value we are predicting is defined as the difference between median earnings of men and women relative to median earnings of men. Our ML problem will receive as an input , vector consisting year, country and selected features, and output gap prediction between the median salaries.

Data Understanding

The data was taken from 2 main sources – world data bank, OECD statistics. Both of them contain many null values that were treated in data preparation state.

World Data Bank – Gender Statistics (GenderStat_Data.csv)

Data is on key gender topics. Themes included are demographics, education, health, labor force, and political participation. Columns are : **Country Name, Country Code (**unique country code), **Indicator Name** (Features topic's were in education, financial, demographic, Health, and more), **Indicator Code**(unique code for feature for other csv use). Other columns represent the **years 1960-2015** (each year has a column).Each row represent measure of a feature in a specific country during the years 1960-2016. Values type in each indicator is float. There are 689 features (indicators) and 263 countries.

OECD (Payment_gap_oecd.csv)

The columns of these csv were:
LOCATION (country code), **INDICATOR(**feature name – always "WAGEGAP"), **SUBJECT (**employees type- 'self –employed'\'TOT'), **MEASURE**(type of measure - 'PC_MENWAGE'), **FREQUENCY** (always 'A'), **TIME**(measured year, range 1970-2015), **Value**(payment gap type float).There are 36 countries in the OECD statistics. We noticed that the wage gap actually decreases throughout the years, and is distributed approximately normal (the year dependency is illustrated at the notebook later on in the prediction part).**both csv's** :

Wage Gap distribution[cell 10]

In order to see specify the wage gap over the years we output wage gap histogram. We can see that the wage gap is decreasing during the years and notice it's normal distribution.

Data Preparation & Cleaning

Csv adjustment for merging

1. Travers years and indicator in Gender stats - **[Cell 3] -** First we removed the columns 'Country Code' and 'Indicator Code' from the data frame. In order to match to OECD we decided to change the country name column to 'country-full-name' in both csv's. We use pivot_table function in order to traverse the 'indicator name' , and year cols.
2. Treating in OECD data – [cell 4] - Adding manually the column 'country-full-name' and defining columns name. Remove duplicates, 'subject' column has caused us duplications in wageGap statistics, so we chose to insert only rows with wage gap statistics on all of the workers and deleted 'self-employed' rows. This choice was made because the majority of the statistics were on 'total' values.
3. Merge both data frames- [cell 5]
   The merge was inner join between the two data frame on 'country-full-name' and 'years' columns in order to intersect between them.

4. Removing meaningless columns –[cell 6]- removed columns with single value-subject, indicator, location, frequency, measure. We also remove in this cell rows with more than 90% null values.

Specify Data types [cells 7-8]

First, we flattened 'country-full-name' column to different categorical columns with get dummies. Each country will by in different column and country value will be binary (treated as int), 0 if this row isn't relevant for specific country else 1. Other columns will be float type.

Missing values [cell 6,9]

First we deleted features with less than 90% valid values. Different thresholds were tested after a few trials, this one was chosen because of common sense and the number of features remaining. In order to fill the rest of the missing values we Perform forward and backward fill. after those 2 stages we were left with 104 features, that 28 of them are binary country indicators.

Train-Test split and Normalization [cells 11-13]

Since the latter years contained much more rows, we took for the test data, all the data starting from the year of 2010. The data is spited to x_test, x_train (all data except wageGap, when test contain year 2010 and above), and y_test,y_train (contains only the wage gap column). Normalization was performed only on x matrix, : to maximize the country influence, we wanted all variables to be in the range of [0,1]. Therefore, we chose the normalization method of $\frac{value - min(value)}{\max(value) - min(value)}$. The whole "year" range will come into account in the formula's denominator(1975-2016). Obviously, the correct and formal way is to split the data first, but this way it is simpler and the correctness is preserved.

Data Modeling

Run faster model – in order to avoid the execution time of the lasso model, please make sure the **'run_model'** variable at the first cell is **False. Pay attention yo your python version (python3 mark true)**

linear regression sklearn [cell 14]

We first used linear regression with "Lasso" regularization. the linear regression was chosen for simplicity and because it was the model that was used at the Article mentioned at the end. The "Lasso" regularization has 2 important properties:

1. It nullifies features by their influence on the training error.
2. For 2 highly correlated features, we have no guarantees on their coefficients distribution

The first property is why we chose this method, which gives us a kind of feature selection (removing all the feature with a coefficient of 0). The second one, means that we should look at the correlation between logically related features, and remove/combine features with respect to their t-test result. This is elaborated at the feature engineering part.

The lasso hyper parameter "alpha" was chosen with LOOCV, which was possible because the data is not too large.

After lasso selection we remain with 55 features (instead of 689 + 31 countries). 25 of them are country dummy variables.

linear regression with statsmodel

For all of our tests and error calculations, we took the average of our predictions vs the wage gap average. This way, visualization is much simpler, and our prediction variance will be factored with (1/number of countries), with the common assumption of an unbiased prediction.

Since we used the "sklearn" library for the lasso regularization, we got a model which we cannot derive its statistics easily. Therefore, we took all the remaining features and ran simple linear regression with the "statsmodel" library. The resulting coefficients are not exactly the same but they are quite similar. For proof of similarity we added a blue line to the graph in the "Model evaluation" part. If you look closely you will find the blue line "peeking" behind the green at minor points.

Statistics & Data Visualization and Engineering

We then started manually engineer features in 3 steps:

1. For comfort reasons, The 'year' feature was added even though the lasso removed it.

2. As stated in the lasso explanation, logically related features correlation was tested and corresponding features were merged/removed.

3. Different attempts for feature modifications, with judgment with respect to the "pearson correlation coefficient".

Correlations [cell 20]

From the remaining features we decided to organize features that are related to separated groups and create correlation matrix between each group to wageGap feature in order to check correlation to wageGap and also to find linear correlation between similar features. You can see the groups in cell 20. Interesting outcomes and further elaboration can be found on the next part of "Correlation Engineering".

Cross-terms correlation [cell 21]

We first took all our remaining features except the country dummy ones. Then, we examined new features with the following modifications:

1. All second order cross terms (which is simply the multiplications between 2 different features, and each feature squared).

2. Taking the square root and the log of the features.

We then tested the absolute value of their "Pearson coefficient" w.r.t the Wage-Gap. This coefficient states how much the two features are linearly correlated, which coincides with our linear regression model. Because we don't care about the sign of it, but its magnitude, we

tested the absolute value. We wanted to add all features with value above 0.5. We got 4 such features, and the fifth one had 0.46 so we decided to take it also.

Engineering By correlations - [cell 22,23,24]

Feature handling was according to the features group correlations and t-test grade after running the modelstats model. You can see the code in cell 22

The features **Age population 0\1\5** has a linear correlation between them. Because the all have a small t-test value (-0.746, 0.020, -0.745) we decided to take the average field of them.

Employment to population ratio features has also a linear correlation between them but because of the high t-test value of each one of them (5.155, 5.943, -6.630, 4.762), we decided to leave them in the features list as is.

The **GDP** features group has also strong linear correlation between GDP and GNI. We removed the futures with the smallest t-test values (GDP per capita-0.117, GNI per capita-0.375).

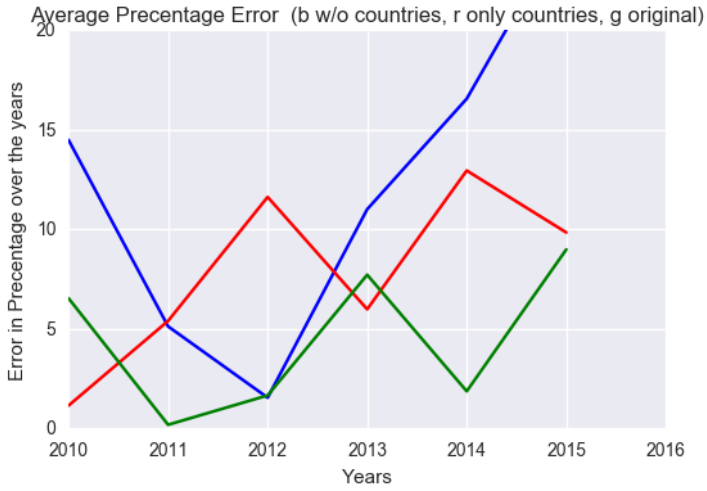
In **Life expectancy** features group, there is also linear correlation between male and females, so we took the average value of both features. We also noticed an exponential correlation between **Mortality rates** to Life expectancy, we decided to leave this feature as is because of the high t-test value - 7.574. we reduced the features from 32 to 26 (without countries).
In cell 23 We also remove low t-test features like 'country-full-name_Spain' and 'GDP'. So overall we remain with total 47 features (and afterwards added extra 5 features in the second method of engineering).

Model Evaluation

1. **countries wage gap through time [cell 15.1]**
   We differentiated between the countries that our model includeded (in green) and the one it does not (in red). you can also see the average wage gap value in blue. 25 countries were taken to the feature list, without an observable context.
2. **Average Wage gap through years [cell 15.2 and 16.1]**
   In order to check our model we plot the average wage gap during the years we had in the test vector and compare is to the average wage gap we predicted with sklearn in cell 15.2. We also verify our prediction against all train and test in cell 16.1. as you can see in the notebook the prediction vs real data is very similar for both test and train.
3. **Average Percentage error on test data vs prediction [cell 15.3 and 16.2]**
   Here we can see the average error in percentage between the test prediction vs the test real data. The error in test is 4.41% and the max error is 9.9%. in addition we calculate the percentage error with all train data on all years (right), which is 2.49% (avg) and 9.9% (max):
4. **Average Wage gap through years and error with statsmodel [cell 18.1, 19.1]**

we plot the average wageGap of predicted data vs test vector and train+test vector comparing the execution of statsmodel and sklearn – we can see very similar results.

5. **Average Wage gap through years real data vs predicted and error in the new model (after changing the features).**
   After executing the linear regression with stats model on the new features list (after engineering and reduction). The results are better but unfortunately the improvement is not significant.
6. **Beating the benchmark [Cell 29]**
   Since our goal is not to predict, but to find irrelevant features, we now want to see how much impact is there on the "country" dummy variables. To do so, we compare 2 models: the first, consists only the our models features without the country variables **[cell 29.1]**, and the other consists solely with country variables and year feature. We should notice that it consists of all 28 country variables and not just those that remained after the lasso. To keep track of previous results we add our previous model also.[cell 29.2.2]




```
average error (without countries) in precentage: 12.34
max error (without countries) in precentage: 25.29
average error (only countries) in precentage: 7.83
max error (only countries) in precentage: 12.95
average error in precentage our model: 2.50
max error in precentage our model: 8.99
```

   As expected, the original model (which includes all the features, and some of the country variables has the lowest error. Unfortunately, our "simply features" model did not beat the benchmark "only countries" model. This could have various reasons, but the important thing to learn from this is that our selected features have different impact at different countries.
7. **Testing by countries [Cell 29.2]**
   In continuation of the "country difference impact" we now change are test-train split. We take each country as test and use all others as train. We then use the model without the countries on it. The average error is measured for each country and compared to

the average distance from wage gap average (at each year). the distance from the average and our model error are expected to have high correlation. From the plotted table we can see nice, but not deterministic correlation. Possible reasons for it is the changing number of samples for each country, and that the countries with the lower gaps are just harder to predict "percentage-wise" .

Conclusions [cell 26.1]

once we got our final model, we generated our final statistics.

The model's average error on the test set is 4.49%. we can see visually from the previous graph that the prediction is quite similar, though the total wage value descends through time so it is harder to get closer by percentage. In addition, the Adj-R^2 is 0.994 which means we truly "captured" most of the models variance.

As stated before, the lasso part removed the 'year' feature even though we clearly see the time dependency. this could be explained (and examined in next section), by the other feature dependency in the years.

We now look for the feature t-test values, and try to make proper inference about the substantial features (P>|t| = 0). Putting our engineered and country features aside, we can see that the most important features include subjects as Employment-Ratio (total, male, and female) and GNI, which corresponds to our initial guess. The most Surprising features (in our opinion) were "Mortality infant" and "Adolescent fertility". besides plotting them, we looked back at those features at the OECD website, and found out why the lasso removed the year dependency (added at appendix).

"Mortality infant" - We presume that these deaths affect the women more than the men and indeed we see that its coefficient is positive (23.3923). If our project needs to "advise" the governments in wage gap reduction, then besides the obvious ways, it might be helpful to somehow financially help women after such cases.

"Adolescent fertility" (births per 1,000 women ages 15-19) - We presume that this is a misleading but important feature. Its coefficient is negative (-14.1793) so it supposes to reduce the wage gap, but our logic says otherwise. A good explanation for it is that those women are totally removed from the "work cycle" by unemployment or by working in jobs who pay cash (cleaning, prostitution, etc.).

As for our evaluations, we can see that the features correspond differently to each country. Though it would be nice to claim that our research produced conclusive outcome, we humbly admit this is not the case.

Future prediction [cell 30]

To test the time dependency, we predicted the average wage gap of 2016 in 2 methods:

1. For each state, we predicted its next feature values based on the past one. Then predicted by the new feature vectors, and averaged the results.

2. Using previous values of the wage gap.

Obviously, the result of such an experiment is just speculation, and deeply relies on the way we predict the future values from the past ones (which we will define as "interpolating").

we chose 2 ways for interpolating: Splines of first order (which is just linear) and second order. For each method (or at least) We hope to get small, but significant difference between the 2 ways, otherwise you could just predict the next wage gap from the past and out model will have no influence. On the other hand, if the difference would be too large, we will get a contradiction to our conclusion of time dependency on the features.

The results:

Spline order	Method 1	Method 2
1	17.95	12.08
2	17.49	16.55

First notice that indeed, our feature prediction does suffer from much less variance, because it averages over the number of the countries.

For spline order 1: The difference of about 5 which is quite substial, which corresponds with the necessity of our model, but the time dependency (meaning the correlation between the values in this case) can still be seen.

For spline order 2: we can see that the results of the 2 values got really close. The best way to explain this, is by pure luck. the extra order seem to correct both predictions towards each other. This empowers the fact that this part is just speculation, and with the true future features is pointless.

So, in the following year, we encourage the reader to come back to this report and check the true wage gap against the methods (with spline order 1) and see which is better :)

External Sources

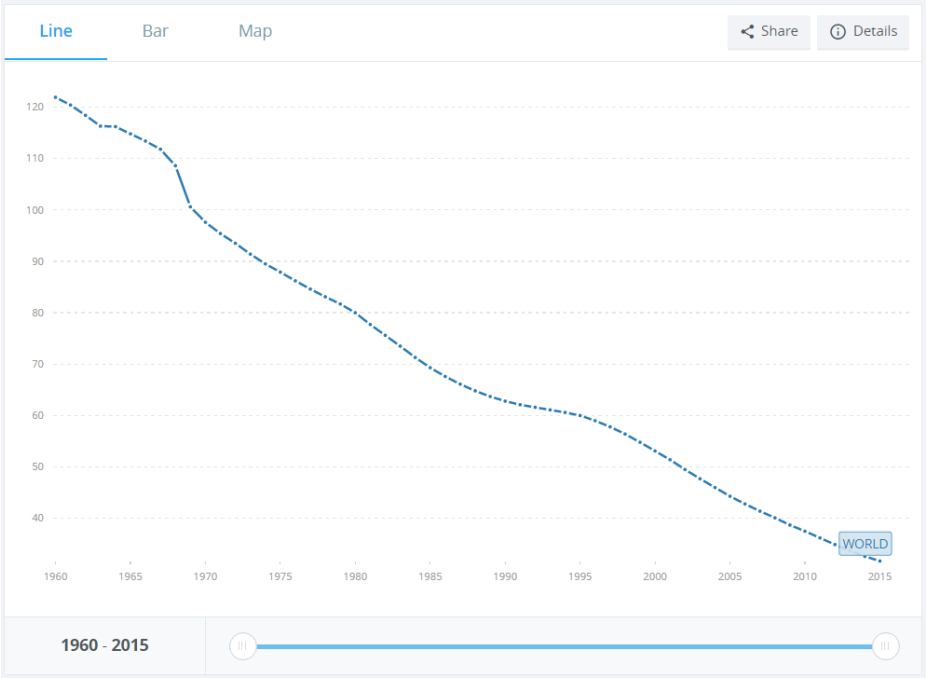
We should mention that throughout this work we used the article "A meta-analysis of the international gender wage gap" by Doris Weichselbaumer & Rudolf Winter-Ebmer, which can be found at http://www.economics.uni-linz.ac.at/papers/2003/wp0311.pdf

Appendix - "Mortality infant" & "Adolscent fertility" graphs from the OECD website

Mortality rate, infant (per 1,000 live births)

Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation (UNICEF, WHO, World Bank, UN DESA Population Division) at childmortality.org. Projected data are from the United Nations Population Division's World Population Prospects; and may in some cases not be consistent with data before the current year.

License: Open



Adolescent fertility rate (births per 1,000 women ages 15-19)

United Nations Population Division, World Population Prospects.

License: Open