**Intro**

Suicides remain the main cause of increased mortality around the world. For governments of different countries it is essential to predict and prevent or, at least, minimize suicides by revealing causes and conditions of this phenomenon.

This research project is inspired by minimizing suicide rate around the world after considering influencing socio-economics factors.

To keep briefness I will use suicide rate instead of numbers of suicides per 100.000 people of population. It’s our target variable, which we want to predict.

The objective of building a predictive model was stated.

**Datasets**

Let me begin by describing used data.

Three datasets were used. Within which “suicide rates overview” is the main one and others consist of supporting data, what is used in hope to increase a quality of predictive model.

Each row in the main dataset contains information about the representative sample of people of number of approximately 30.000

Mental disorders rates are presented as the percentages of people who are struggling from these diseases in a particular country and year

As we can see the countries in our data are presented by western part of the world. Also there are some Central-Asian countries, countries from Oceania and Africa.

**Describing analysis**

Let’s turn to describing analysis.

There are big outliers in our data up to 500 points (500 people per 100.000 of population)

In order to prevent difficulties with teaching a stable model only data below 95th percentile was used in building a model.

Our data is full only in margins between 1985 and 2016 years

If we consider top 10 countries by suicide rate from recent years we can see Lithuania as antileader in this rating. In the middle of 90-s the half of first 10 consists of post-soviet and Eastern European countries - consequence of demographic crisis.

There is no constant trend of suicide rate in marging from 1985 to 2016.

Local peak is reached in the half of 90-s - consequence of demographic crisis in post-soviet countries and eastern Europe

Approximately 77 percents of suicides belongs to the male.

Middle aged people usually commit 36% of all suicides.

After age and gender grouping we can see that in general men are more likely to commit suicide, especially men between 34 and 54. It can be a consequence of hard-working style of living and unfair division of labor between women and men.

**Hypotheses**

Let’s consider the stated hypothesis.

In general hypothesis about statistically significant interactions between age, gender, generation, gdp\_per\_capita, country income level, HDI, unemployment, inflation and suicide rate was stated.

All hypotheses concerning social classification were proved, as well as hypothesis about country income grouping. That’s why categorical parameters as gender, generation, age, country income were included in the final predictive model.

Some mental disorders rates can be used for accurate suicide rate prediction.

**Methods**

Let’s consider methodics that were used in analysis

There are three main groups of hypothesis were stated.

* Influence of classification:
  + - Application of statistical criteria for group equality, evaluation of data distribution.

In general our data is not normally distributed that is why instead of Tukey comparison criterion U-criterium was applied.

Boxplot chart allows us to compare groups outliers and make a conclusion regarding the variance in our data.

* Numeric interconnection:
  + - Scatterplot, correlation coefficient calculation, linear regression

For revealing interconnection between metric variables linear regression model was applied as well as scatterplot and correlation coefficient calculation

Linear methods of analysis provide us an ability to reveal valuable metric predictors that can be used in a predictive model.

* Time series analysis:
  + - Trend margin framing, linear regression

During analysis of time series trend margin framing and trendline (if it was possible) were applied. After that explanation concerning crises and other extreme situations was introduced.

**Essential observations**

1. As was mentioned earlier in general men are more likely to commit suicide, especially men between 35 and 54. It can be a consequence of hard-working style of living and unfair division of labor between women and men.
2. We can see a tendency of suicide rate decreasing for each younger generation. Generation Z (in fact my generation) is exposed to committing suicide least of all comparing with other generation groups. Representatives of G. I. and Silent Generation show the highest suicide rate, which may be due to the difficult events of most of their lives (2 world wars, great depression, revolutions in a number of countries, pandemics)

On this slide you can see corresponding years of birth for each generation group

Generation Z - 1997-2012

Millennials - 1981 - 96

Generation X - 1965-1980

Boomers - 1946-64

Silent - 1928-45

G.I. Generation - 1901-27

1. Remarkably, including alcohol and depression rates in linear regression predicts 33% of variance in our data. These predictors were included in the final predictive model. On the other hand, including GDP per capita in linear regression shows worse results and multicollinearity. This metric was not included in the final model.
2. Surprisingly, HDI is not negatively interconnected with suicide rate but positively(!!!) The more the HDI in the particular country the more the suicide rate there. However, this interconnection is quite low - correlation coefficient is 0.1. It can be explained with the fact that the richer the country the better it leads statistics on suicides. The HDI was not included in the final predictive model.
3. According to time series and income level grouping analysis countries differ from each other. The variable “incomelevel” was included in the final predictive model. As we can see, countries with high income in average present the biggest suicide rate. High-income countries have higher suicide rates than low-income countries for almost all time intervals except 1994 to 1996 and 2014 to 2016. Countries with low incomes show a very strong peak in the suicide rate for 1993-1995, which may be a consequence of the crisis of the 90s in the post-Soviet and Eastern European countries.
4. After interconnection analysis between unemployment, inflation and target variable was held it turned out that these economic indicators have quite low degree of linear interconnection with suicide rate. These variables were not included in the final predictive model.

**Building a model**

* After essential socio-economic factors were revealed several intermediate predictive models were built with different numbers of predictors.
* As the trained model decision tree regressor was used from the python package “sklearn”.
* In order to increase the prediction quality of models and to decrease the probability of overfitting data lower than 95th percentile was utilized.
* During cross-validation as the metric of quality RMSLE (root mean squared logarithmic error) was applied to dispose of the effect of outliers.
* Intermediate models were compared within each other using RMSE (the root of the mean square of the error - classic metric for observing the absolute error) and R ^ 2 (the coefficient of determination - the percentage of explained variance)

**Cross-validation**

* The training itself took place using HalvingSearchCV cross-validation to divide into groups to prevent overfitting.
* On each iteration training data is divided in k parts, where k-1 parts are used for training and the last - for validating model metrics.
* HalvingSearchCV allows to train several models and save the one with the best value of target metric (RMSLE in our case) and optimal hyperparameters:

depth, min\_sample\_split, min\_sample\_leaf

The search strategy starts evaluating all the candidates with a small amount of resources (as time and number of calculations) and iteratively selects specified part of the best candidates, using more and more resources.

General steps in training each model

The depth of default tree was used as upper bound, min-samples-leaf and min-samples-split were permuted from the set with the step of hundred points

**Conclusion on model building**

The model with age, generation, gender, level of prosperity of the country, level of alcoholism and depressive disorders showed the best values of RMSE and R^2 metrics. It can be used for predicting suicide rate with absolute error of 1.1 points what is quite precise.

I want to draw your attention to the point that data on psychiatric disorders provided the largest contribution to the training of our model, confirming their significant impact on the suicide rate.

In the end, I would like to recall the key points of the study and interesting observations.

**Summarizing**

1. Project goals were achieved:
   1. Statistically significant socio-economic factors have been found that affect the rate of suicide (per 100,000 population)
   2. A model has been built that predicts the value of the target variable with high accuracy
2. Interesting observations made:
   1. Men are the most risky group.
   2. Younger social groups have a lower risk of suicide.
   3. Rates of psychiatric abnormalities help improve the prognosis of the suicide rate.
3. A basis for further research has been obtained: a detailed analysis of each of the divisions of observations is possible - by age, gender and generations.

Well, that brings me to the end of my speech and I will be glad to answer your questions. Thank you for your attention.