

FICO™

# Exploratory data analysis with Pandas

## [mlcourse.ai](https://mlcourse.ai) –Machine Learning Course

Yuri Chekalin

Open Data Science Community (<https://ods.ai/en>)

# Agenda

## Exploratory data analysis with Pandas

- ▶ EDA: investigating the data
- ▶ Churn prediction problem in Telecoms
- ▶ NumPy and Pandas data types
- ▶ Main Pandas features (Jupyter Notebook)
- ▶ Building prediction model (Jupyter Notebook)
- ▶ Data cleaning concepts

Download this lecture at  
[https://github.com/DmitriiDenisov/mlcourse\\_dubai](https://github.com/DmitriiDenisov/mlcourse_dubai)

# EDA: investigating the data



- ▶ Don't underestimate data exploration!
- ▶ What means "to know your data":
  - ▶ Dataset size and variable types
  - ▶ Distributions of variables
  - ▶ Noise level (how clean is the data)
  - ▶ Predictive power of variables and correlations
- ▶ EDA will allow you to plan next steps
- ▶ Sometimes investigation results show that data is simply not good enough
- ▶ EDA can be based on numbers or visuals
- ▶ EDA helps in model reporting

# Churn prediction problem in Telecoms



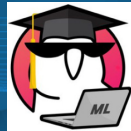
- ▶ “Churn” can be defined differently, this need to be agreed in advance
  - ▶ voluntary attrition, operator switch
  - ▶ silent churn, not using account
  - ▶ case of labelled data
- ▶ Churn prediction goals:
  - ▶ Business goal is skipped (but normally – to decrease the churn)
  - ▶ Technical goal - to explore the data, get insights and build prediction model
- ▶ Given dataset parameters
  - ▶ Data is cleaned and ready to be used in prediction (never expect this in real life)
  - ▶ “Churn” is defined (labelled data)

# Basic data categories in data science



Variables predictive science perspective:

- ▶ Numeric (Continuous) - Type Float, Integer
- ▶ Categorical - String
- ▶ Ordinal - String, Integer
- ▶ Binary - Boolean, Integer, String
- ▶ Date/time
- ▶ Target – any (depending on the problem)



► Let's go coding!

# Data cleaning concepts: bad data types



Bad data types:

- ▶ Missing values
- ▶ Irregular data (outliers)
- ▶ Skewness (not Normal distribution)
- ▶ Unnecessary data
- ▶ Inconsistent data



# Data cleaning concepts: Missing values



Missing data:

- ▶ Naming: empty value, missing value, missings, Null, NaN
- ▶ If missings have different nature – they need to be marked
- ▶ Zero values can be ‘masked’ missings
- ▶ Some model types can not work with missings
- ▶ What to do with missings
  - ▶ Delete columns with missings
  - ▶ Delete rows with missings
  - ▶ Impute missings (ex. Replace by average values)
  - ▶ Replace missings (ex. -999, ‘\_MISSING\_001’)



# Data cleaning concepts: Outliers



- ▶ Easy to find (standard plots and functions)
- ▶ Should be treated based upon the problem, dataset and the project goal
- ▶ Sometimes outliers is what you actually need! Examples:
  - ▶ Payment fraud detection
  - ▶ Network security breach detection

# Data cleaning concepts: Skewness



- ▶ Normally distributed variables are better predictors
- ▶ Logarithmic transformation  $y = \log(x)$  often helps

# Data cleaning concepts: Repetitions & duplicates



- ▶ Repetitions require further investigation within data source
- ▶ Possibilities
  - ▶ Unnecessary characteristic
  - ▶ Top #1 predictor
  - ▶ Basis of segmentation model
- ▶ Duplicated columns should be deleted, but double check before!
- ▶ Duplicated rows should be investigated
  - ▶ Sometimes you will insert duplicated rows by yourself!

# Data cleaning concepts: Inconsistent data



- ▶ Capitalization ('Bad', 'BAD', 'bad') – to be lowercased
- ▶ Wrong data formats – find and correct
- ▶ Wrong encoding for categorical vars:
  - ▶ Ex. Gender can be: 'M/F', 'Male/Female', '0/1', '1/0', '0/1/2')
- ▶ Addresses encoded in one string

# Resources used in this lecture



- ▶ MLCourse.AI lecture #1:  
<https://mlcourse.ai/articles/topic1-exploratory-data-analysis-with-pandas/>
- ▶ Notebook "Comprehensive data exploration with Python":  
<https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python#5.-Getting-hard-core>
- ▶ Kaggle competition "House Prices: Advanced Regression Techniques":  
<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
- ▶ Article "Data Cleaning in Python: the Ultimate Guide (2020)":  
<https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d>

# Thank You

**Yuri Chekalin**

YChekalin@gmail.com

Mob. 054 441 6388

WhatsApp: +7 985 226 6049



**Yuri Chekalin**

Senior Data Scientist & Pre-Sales Consultant



**FICO**<sup>TM</sup>

