Project 2 - Data Science Bootcamp

Final Deliverables:

1. Slide deck PDF pushed to GitHub designed for non-technical stakeholders- that outline findings and recommendations, and future work (10min presentation).
2. Jupyter notebook following PEP8 designed for data science/ technical audience.

Things to think about

- Try different (at least 3) machine learning algorithms to check which performs best on the problem at hand
- What would be the right performance metric- precision, recall, accuracy, F1 score, or something else? (Check TPR?)
- Check for Data imbalance

Time until: Thursday 08.10.2020 11:00

1. Lending Club

For this project we chose a dataset from Lending Club **approved personal loans between 2007 and 2011**. The data can be found on www.lendingclub.com. The purpose of the analysis is to reduce defaults, improve profitability and help the company and investors determine interest rates. We will use machine learning models to analyze credit risk as a binary classification problem.

How to choose the Performance metrics? - well the model (whichever you pick) will be used to determine who should be approved for a loan and who shouldn't, denying the loan to a client who will end up paying in full (false positives) represents a loss, but because interest is usually only a portion of principal the company will most likely be more comfortable not taking the chance when the risk is not to get reimbursed at all and lose the entire principal which represents a higher amount. Thus the main concern here is to avoid approving somebody who won't be able to repay or in other words avoid false negatives. This is achieved by a model with a high recall rate.
What would be the right performance metric- precision, recall, accuracy, F1 score, or something else?

We also might need to evaluate TPR to make sure we are not declined too many qualified borrowers.

Make sure to check data imbalance.

2. Home credit default

Home Credit is a lender who provides loans to populations unable to use traditional credit services.
Goal is to lower loan risk by identifying patterns from within historical data.
Let's figure out the most predictive data points through some machine learning models.

3. Kickstarter Project Success

In recent years, the range of funding options for projects created by individuals and small companies has expanded considerably. In addition to savings, bank loans, friends & family

funding and other traditional options, crowdfunding has become a popular and readily available alternative.

Kickstarter, founded in 2009, is one particularly well-known and popular crowdfunding platform. It has an all-or-nothing funding model, whereby a project is only funded if it meets its goal amount; otherwise no money is given by backers to a project.
A huge variety of factors contribute to the success or failure of a project — in general, and also on Kickstarter. Some of these are able to be quantified or categorized, which allows for the construction of a model to attempt to predict whether a project will succeed or not. The aim of this project is to construct such a model and also to analyse Kickstarter project data more generally, in order to help potential project creators assess whether or not Kickstarter is a good funding option for them, and what their chances of success are.

4. Personality vs. Risk of Drug Use

The abuse of tobacco, alcohol and illicit drugs is costly to our society. According to the National Institute on Drug Abuse, more than 740 billion is lost annually due to lost work productivity, crime and healthcare.

People with certain personality traits may be at increased risk for drug use problems, and studying personality may help researchers better understand and treat these problems. **Attempting to link genetic disposition to substance use disorders has proven unsuccessful.** There are however many studies that have shown positive correlations between risky personalities and drug use. Understanding this relation could lead to better treatment.

The dataset used for this project is referenced. There are limitations to the study since the collected sample is biased with respect to the general world population but it remains useful for risk evaluation. Each column of attributes has been normalized using T-scores and is described below.

T Score
The term **t score** has different meanings in different settings:

- In introductory statistics, t score is often used synonymously with t statistic.
- In psychometrics, a t score is a type of standard score computed by multiplying a z-score by 10 and adding 50.
- In bone density tests, a t score compares bone mineral density to a reference mean.

Often, the meaning is clear from the context. In educational research, the psychometric definition generally applies. In papers on osteoporosis, the bone density definition is probably the intended definition.

Sometimes, though, the meaning is not clear. Some statistics texts use the term to mean "t statistic". Others use the psychometric definition.
To avoid confusion, the Stat Trek website avoids the term "t score". It uses "t statistic", instead.]

EScore: Escore (Real) is NEO-FFI-R Extraversion. Extraversion is one of the five personality traits of the Big Five personality theory. It indicates how outgoing and social a person is. A person who scores high in extraversion on a personality test is the life of the party.

Might have to figure out the other kinds of scores from the internet.

Some references:

1. Drug, Wikipedia URL: https://en.wikipedia.org/wiki/Drug
2. The Five Factor Model of personality Model of Personality and Evaluation of Drug Consumption risk, E.Fehrman, A.K. Muhammad, E.M. Mirkes, V. Egan, A.N Gorban. URL: https://arxiv.org/abs/1506.06297
3. Detecting and Assessing Alcohol and Other Drug Use. URL: https://www.ncbi.nlm.nih.gov/books/NBK236259/
4. Ibid.
5. UCI-Machine Learning Repository URL: archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
6. Numpy Documentation
7. Matplotlib Documentation
8. Seaborn Documentation
9. Pandas Documentation.

Lending Club	**Home Credit Default**	**Kickstarter**	**Drug Risk**
Silas/Jonas		**Felix/Tjade**	**Mirko/Nina**
Andre/Leo		**JJ/Olaf**	**David/Niko**
		Jacqueline/Tobi	
		Raphaela/Britta	

Gruppe:			Time slot on Monday
Gruppe 1	André	Leonard	13:20-13:40
Gruppe 2	Britta	Raphaela	14:00-14:20
Gruppe 3	David	Niko	14:20-14:40
Gruppe 4	Jacqueline	Tobi	15:00-15:20
Gruppe 5	JJ	Olaf	15:20-15:40
Gruppe 6	Jonas	Silas	16:00-16:20
Gruppe 7	Mirko	Nina	16:20-16:40
Gruppe 8	Felix	Tjade	16:40-17:00

Gruppe:			Presentation Time
Gruppe 1	André	Leonard	11:00-11:20
Gruppe 2	Britta	Raphaela	11:20-11:40
Gruppe 3	David	Niko	11:40-12:00
Gruppe 4	Jacqueline	Tobi	13:00-13:20
Gruppe 5	JJ	Olaf	13:20-14:00
Gruppe 6	Jonas	Silas	14:00-14:20
Gruppe 7	Mirko	Nina	14:30-14:50
Gruppe 8	Felix	Tjade	14:50-15:10