**Call 25/03/2021**

**Prof. Przanowski**

1. Access to real data: Research for real data.
2. Randomly generated data

**Proposals**

* Comapring and explaining ML algorithms
* Creation of data: build customer characteristics based on manual work. ABT (analytical based table).

**Encoders**

Trying to predict variables.

Convolutional neural networks -> 7.1 Learned Features

We don’t create customer characteristics by myself but create data based on neural network and then we can build models.

* Transform codes from SAS to Python

Call 26/03/2021

Think what sort of problem I’d like to work on. **Two fields to work on:**

1. Credit-risk management.
2. Market-risk management.

<https://www.kaggle.com/c/home-credit-default-risk>

**3 subjects to think on for the dissertation:**

Data contains data rergarding gender (female and amle), multidimensional (many variables),

In scoring the are **3** main datasets = **application scoring** (filling a form in the bank), **behavioural scoring** (bank knows overdues, payments, etc), **external datasets** (credit bureaus, the bank checks with a credit beaureu if I have a credit in another bank).

**Mostly use Python.**

1. Performance **measure** focus: focusing on best model taking training data set and predit target variable. Measure of best fit Using gini coefficient (preferrable).
2. **Explain** the models logistic regression is easy to interpret, XGBoost/neural network/random forest is hard to interpret (explain what’s going on in these models).
3. **Fare aware modelling**  the model shoudnlt be biased by gender, religion, skin complexion, etc (sensirtive variable), these variables by default in the USA are not used on modelling to avoid problems.

**Harvard business reivew: when machine learning goes off the rails BABIC Jan 2021**

**02/07 meerting**

**Book: The law of equal opportunities or unintended consequences? -> Woman are better payers of the loans**

Fred Pambel

Chapter 1: Credit Scoring

* best practises
* classical scoring approach
* drawbacks of those classical scoring approaches
* some obstacles present in those.
* Economics on credit scoring: **check przanowski book**

Chapter 2:

Explain how the previous drawbacks are solved with ML automatically (not manually)

* Logistic regression vs other ML models
* Problem of multicollinearity
* Omitted variables problem
* Negative betas

Chapter 3:

* Additional obstacles/complications related to credit scoring framework: regulations, ethics, etc. Giving a credit on a model is just a 0 or 1, but for a person is the capacity to buy a house or not. We need governance
* [Human oversight to the explainability of the model](https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html)
* Fair aware problem
* Omitted variable problem: what will happen if we delete variables. (**ADD 2 PAPERS ask Daniel)**

Chapter 4:

* Numerical example with Kaggle data
* EDA, estimation of model, compare variance of models, bla bla.

Chapter 5: Reverse engineering from app <https://github.com/sachinrajput17/Loan_Prediction>

matrix decomposition cholesky

To add to the drawbacks of those classical approaches subchapter:

* Explain what is “weight of evidence” (WOE) transformation and “information value” (IV) and provide an example. Why these techniques are important?
* Explain that on classical approach we are not able to model nonlinear relationships between variables.
* problem with selecting features: Explain that variables cannot be selected automatically using statistical tests or Machine learning techniques, they must be done manually.
* Explain that there are not interaction between variables in classical approach: if we observe negative betas we must then create an additional feature (beta1\*beta2) manually that represent this “nonlinear” relationship between variables.
* Explain that many of these things that are done manually in classical approach, can be also done automatically with machine learning. And slightly give an explanation to what is machine learning and its applications in the fintech sector, especially in credit scoring.

**To add to the Logistic regression vs other ML models chapter:**

* How to transform from linear regression to logistic regression (from the book of Fred Pambel), assumptions, transformations, interpretation of coefficients in linear vs logistic regression, advantages, disadvantages.

3 layers of interpretation in logistic regression (depending on how the equiation is presented):

* Probability
* Odds
* Logit

**To add to the Problem of multicollinearity chapter:**

Lets asumme we know the exact underlying process of generating data (real equation) Y = B0 + B1 + Bn

Then we delete 1 important variable, for example: **sex** that we know in advanced that that variable affects the credit score), we then observe a **bias** on the left parameters that are left. What happen to the parameters that are left after the removal?

How to conduct the explanation?.

* Create a model that contains this relevant variable
* We assume that this feature is important, then the model performance will drop
* When this relevant variable is correlated with other attributes, when removing this variable, the information/impact that is included in this sensitive variable is transmitted to other variables to which it is correlated.
* It won’t be observed a drop in accuracy of the model since the other variables absorved the impact of the deleted important variable that was correlated. We will however observe an increase in bias. **WE WILL HAVE A BIASED MODEL**

The problem is that the regulator requires us to provide information to the customer with a biased model.

Transform this explanation to a empirical example (from Kaggle)

**To add to the Negative betas problem sub chapter:**

If we create logistic regression based on WOE (weight of evidence), IF WE OBSERVE NEGATIVE BETAS, it means that we have sympson’s paradox in our data (higher other interaction between variables). IF WE HAVE POSITIVE BETAS, it means that we don’t have any interaction between features.

If we have negative betas but WOE was positive slope it means that there are some interaction between features. (Check prof Kaszynski book)