Leveraging data science & machine learning to enable better credit decisions

Plan

1. Identifying use cases in the credit industry

2. Example of credit scoring modelling

1.

Identifying use cases in the credit industry(*)

Granting new credits

Developing credit scoring models could enable

Better user experience to increase conversion rate on our online platform

Focus on next slide

- ❖ Automatically granting loans to customers most likely to pay back
- ❖ Proposing custom loans (amount x rate x term) that would be automatically accepted based on each individual prospects' profiles

Better pricing to increase revenues based on the customer risk strategy

Focus on next slide

- ❖ Optimizing prices per risk profile
- ❖ Dynamic pricing to encourage specific risk profiles to subscribe to loans (the more we want a risk profile, the lower the price, and inversely)

Valuing the customer base

Assessing customers current payment profiles with machine learning could enable

Up-selling / cross-selling good customers to increase revenues

- ❖ Customers with good repayment profiles and where we could increase our exposure could be offered additional loans
- ❖ Customers could be offered to repay their loans earlier to increase our capital to reallocate it to more profitable risk profiles

Addressing customers in pain to reduce losses

- ❖ Renegotiating loans for payers past due to avoid default
- ❖ Prioritizing debt collection efforts

Credit industry(*): we focus solely on selling loans and monitoring the value and the risk of a portfolio of loans with different interest rates, maturities, etc. There are many related topics to credit that we don't cover here for simplicity, for instance risk modelling, credit insurance, marketing, fraud detection…

data iku

1.

Identifying use cases in the credit industry(*)

Focus on the use case:
“Automatically granting loans to customers most likely to pay back”

Let’s say we have developed the following models

credit scoring model #1	Estimates the probability ***p*** that a prospect will pay back fully in the future a given credit application.
credit scoring model #2	Estimates the ratio ***R*** of the full amount that will be paid back by a prospect in the future divided by the initial amount granted for a given credit application.

Then we can develop the following heuristics

Business heuristic	If ***p*** >90% and ***R*** >1.20 then automatically grant loan online.

The values of 90% and 1.20 are illustrative here and they have to be analysed with actual data and risk of default and losses have to be estimated.

1.

Identifying use cases in the credit industry(*)

Focus on the use case:
“Optimizing prices per risk profile”

Let’s say we have developed the following model and analysis

credit scoring model	Assesses the risk profile ***w*** (values range from A to G) of a given applicant purely based on applicant data, for instance age, revenues per year, number of bank accounts with credit cards.
Analysis	The repartition of risk profiles for our portfolio of loans is 20% A, 20% B, **30% C**, 20% D, **10% E**.

Then we can develop the following heuristics

Business heuristic	New applicants with a “C” risk profile have an interest rate bumped by 0.3% to reduce our exposure to C profiles. New applicants with “E” risk profiles have an interest rate decreased by 0.5% to increase our exposure to “E” risk profiles.

The values of +0.3% and -0.5%, same with the repartition of risk profiles in the portfolio, are illustrative here and they have to be analysed with actual data and risk of default and losses have to be estimated.

1.

Identifying use cases in the credit industry(*)

In parallel with use cases identification, defining an ambitious data strategy will increase the chances of success of each individual use case

- **Augmenting the volume and the quality of data used by models to get more accurate models**
  - purchasing (alternative) data from data providers (IQVIA, FactSet, …)
  - allowing prospects & customers to share their own data (with banking APIs…)
  - collecting more customers data
- **Using models more aligned with the business strategy**
  - Choosing the right algorithm to have the right balance between accuracy, interpretability
  - Aligning with regulatory requirements, creating documentation
  - Defining the “ML Ops” strategy: governance and repository of models, frequency to retrain (automatically?) with “fresh” data, handling of bugs, data & accuracy drifts monitoring…
- **Empowering the workforce to leverage these models and data the best way**
  - Creating synergies between data and model experts and business experts
  - Testing POCs, encouraging initiatives
  - Using an enterprise data science platform to empower the workforce
  - Providing data science trainings to foster innovation and emulation

2.

Example of credit scoring modelling

Create a credit scoring model to assess the likelihood to be delinquent in the next 2 years

Step 1: Data analysis, cleaning and preparation for “training”
Step 2: Defining the measure of accuracy and testing different modelling approaches to identify the best model for this given dataset
Step 3: Interpreting results and getting model ready for inference

Step 1: data analysis

- **We have enough data to build models** (150k rows is great, minimum: ~1000 rows)
- **We could have more columns** *(10 columns is not much since models can handle ~100s or 1000s of columns)*
- **There are minor non blocking data quality issues**: 20% missing values for MonthlyIncome, 3% for NumberOfDependents

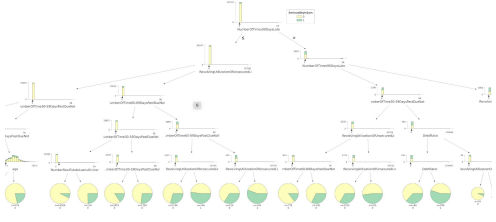
Step 2: model selection

- **Simple models are as good as advanced models** (which is expected with as few as 10 columns)
- **A Decision Tree model fits well with this problem**: straightforward data preparation, simple model (fewer chances to learn weird patterns from the data), highly interpretable, easy to put in production

Step 3: interpretation

Key insights

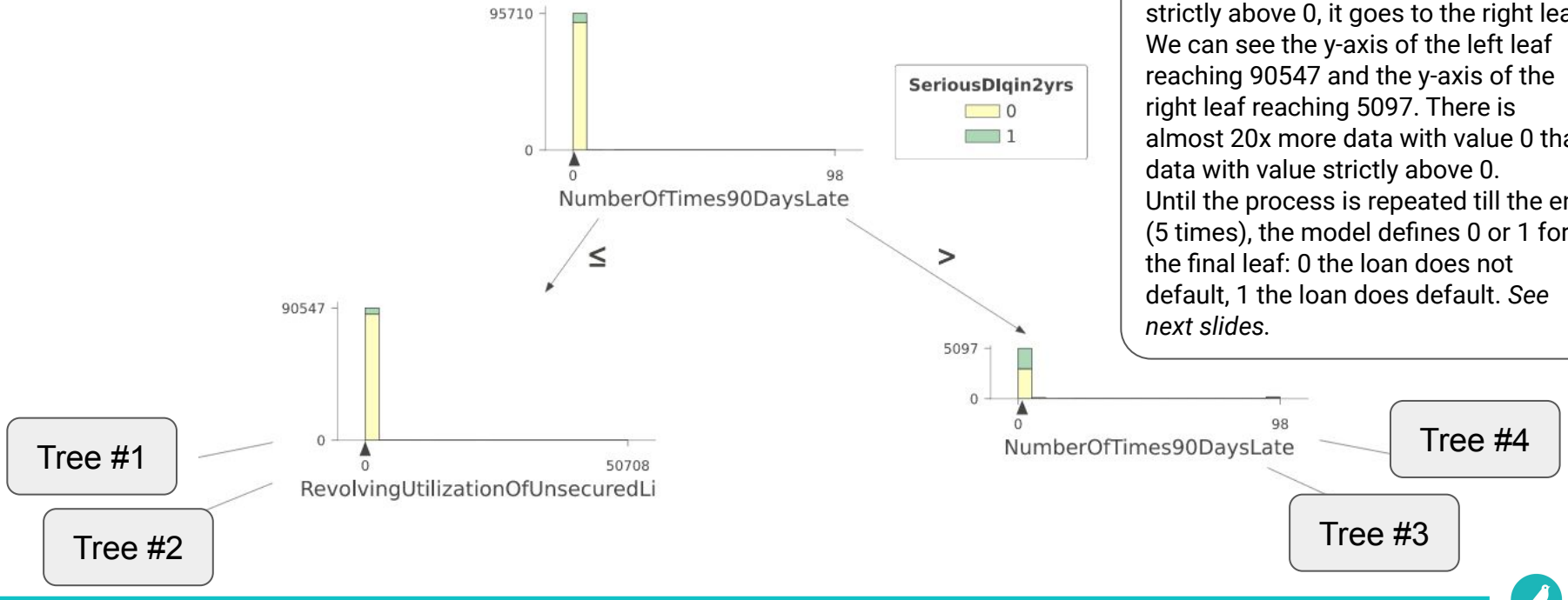
- Most predictive variables: being late on historical payments
- Never late on payment: very safe
- Late more than once: unsafe



2.

Example of credit scoring modelling

Top of the Decision Tree

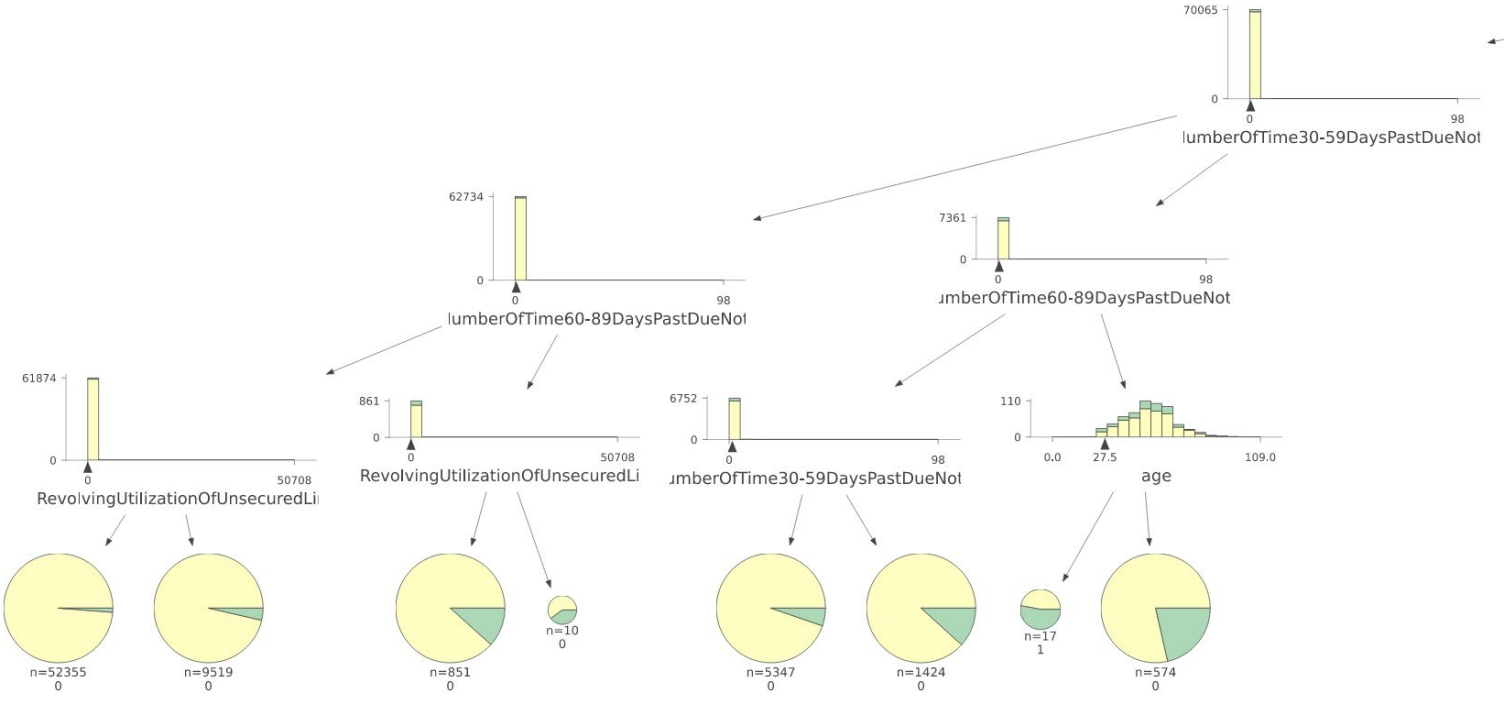


💡**How to read the Decision tree?**
Data is split at each “node” of the tree in 2 “leaves”, always using 1 variable. This process is repeated 5 times (5 was giving the best accuracy for the model). For instance, the first node uses the variable NumberOfTimes90DaysLate to split the training data in 2: if NumberOfTimes90DaysLate is equal to 0 or less, it goes to the left leaf. If it is strictly above 0, it goes to the right leaf. We can see the y-axis of the left leaf reaching 90547 and the y-axis of the right leaf reaching 5097. There is almost 20x more data with value 0 than data with value strictly above 0.
Until the process is repeated till the end (5 times), the model defines 0 or 1 for the final leaf: 0 the loan does not default, 1 the loan does default. *See next slides.*

Example of credit scoring modelling

Bottom of the Decision Tree

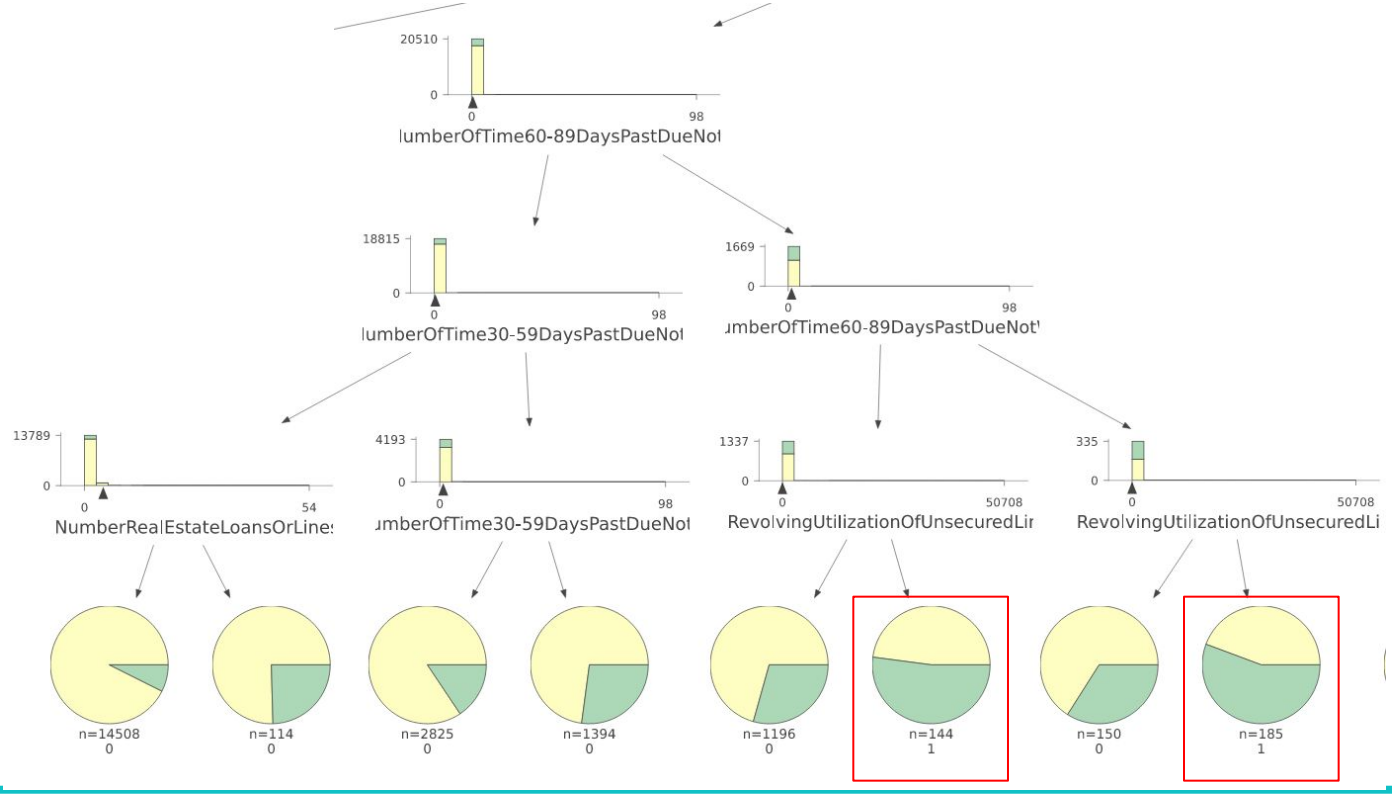
Tree #1



Example of credit scoring modelling

Bottom of the Decision Tree

Tree #2



Example of credit scoring modelling

Bottom of the Decision Tree

Tree #3

2.

Example of credit scoring modelling

Bottom of the Decision Tree

Tree #4