Machine Learning Summer 2022
Imperial College Business School

Final project: Risk Models for Rent Arrears

Wolfram Wiesemann
Mark Van Lokeren

Context

The goal of the project is to devise an efficient, targeted strategy to detect and prioritise tenants at risk for falling into arrears for a rental housing association in the US.

Data

The data is contained in the file `rental_data.csv` and consists of about 56k rental payment transactions recorded by a housing association in the US.

- name = the first and last name of the tenant
- dob = the date the tenant was born
- houseID = a key representing a single rental property
- houseZip = the 5-digit zip code of the rental property (US zip codes)
- paymentAmount = the amount payed in rent on the payment date
- rentAmount = the amount of rent due

Objectives

(a) Can you and your group differentiate between those who are more at risk of entering a state of long-term arrears (which accounts for the majority of arrears debt) vs those who are likely to pay back their debt quickly? To be more specific, who is likely to fall behind on payments for over six months and who is likely to catch up?

(b) Is there a single factor with the strongest predictive power, or is it a subtle combination of factors? How do these combinations vary?

(c) Can you produce a predictive model that produces a score card for the risk of long-term arrears for each tenant?

(d) Can you extend this model to differentiate between short-term and long-term arrears risk?

You should attempt at least **3 different models** and compare their performances. A motivation for your choice of models is essential. Highlight any strengths and weaknesses of each of the models you try. Tabulate your results for easy comparison and discuss any similarities or differences.

For this problem it helps specifying grades of risk, based on for example, how often a person enters in arrears over a period of one year. So, let us say that in a 12-months window,

- *low risk* means one month or less of accumulated arrears,
- *medium risk* means between one and five months of accumulated arrears, and
- *high risk* is to imply that a person has six or more months of accumulated arrears.

Both the window size and the risk grades can be arguably adjusted. In a way, they are also hyperparameters of your model.

Based on your findings from Objectives (b)-(d) above, is it possible to translate your analytical insights into actionable business insights? For instance, can we help the housing association to determine which of the arrears cases pose the greatest risk, and should, therefore, be prioritised? There is a financial cost to the association, and a personal impact on the tenant, and both should be taken into account. Assume that the overall cost of a high risk tenant is about 10 times that of a low risk tenant. Some questions to ponder about are:

(e) How should the housing association deal with medium risk tenants?

(f) Is it possible to help prevent people from falling into long-term arrears?

Deliverables

Your report should be submitted in a **Jupyter notebook** and as a **PDF file**, clearly detailing the steps involved in your analysis and any assumptions you have made along the way. Any additional files you will have used to support your analysis should be submitted too.

Checklist

Before submitting your report, check whether it fulfils the following:

- ☐ presents and summarises the data using descriptive statistics;
- ☐ provides context for the analysis;
- ☐ presents the models used in the analysis together with their main characteristics;
- ☐ uses tables, figures or graphs to enhance the clarity of the information;
- ☐ includes an executive summary;
- ☐ discusses the strengths and limitations of the analysis.

Grade

While you will only receive an overall grade for the coursework (together with written feedback), you can expect this overall grade to be based on three components:

- correctness and completeness of the employed models;
- value of the suggested managerial actions;
- quality of the presentation (e.g., powerful tables/visuals, clarity of reasoning, to-the-point advice, professional layout and absence of typos).

Not contributing to a high grade (just to be clear) are the number of pages written or the inclusion of irrelevant material (including fancy figures/graphics or paragraphs that do not support the discussion).