ML_project_1 - Predicting which businesses are unlikely to pay their fines

Predicting which businesses are unlikely to pay their fines

Using ML to predict which businesses are unlikely to pay their fines issued by the registry department of Tartu County Court (the Estonian Business Register)

The back story

All legal entities in Estonia are register in the Estonian Business Register.
While a business is registered, they have the duty to keep their data up to date in the register and submit an annual fiscal report.
Failing to do either, can result in the business being fined.
There are thousands of fines issued by the Business Register every year.
The levels of fines being payed by legal entities are not good.

The goal

Our goal is to find the legal entities that are unlikely to pay their fines and to do so even before the fine is issued.
There is a real life use case and a successful project could help improve the Estonian business environment.

The data

There are 348 465 legal entities registered in the Estonian Business Register and there have been 34 528 fines issued in 2024 alone (as on 20.09.2024).
Main reason for the fines being issued is the annual fiscal report not being submitted (on time).

The data comes in 3 data tables:

1. ML_P09_1legal_entities
   - the main table to work with
   - contains data on all legal entities in Estonia that were in the register on 01.01.2019 or were registered after that date
2. ML_P09_2phys_persons
   a. The table includes physical persons who have a connection with legal entities
3. ML_P09_3relations
   a. This is the relations table connecting the legal_entities to the physical persons.
   b. One legal_entity can have many connections to physical persons.
   One physical person can have connections to many legal entities.

Further information

To understand the Estonian business register better, check out the e-Business register statistics page (in Estonian and English).

- https://ariregister.rik.ee/eng/statistics

To understand the tax data, check out the EMTA homepage:

- https://www.emta.ee/

A press release from the Justice ministry that might be of interest:

- https://www.just.ee/uudised/majandusaasta-aruanded-esitasid-sel-aastal-tahtajaks-66-kohuslastest

NDA

This is the serious bit.

We are using a lot of data and the tables have been put together specifically for this project.
The data belongs to the Estonian Justice Ministry and can only be used for the purpose of this project.
It is very important that the data is handled properly, following the Estonian law and the best practices of working with data.

Our legal team has put together an NDA and guides to follow when working with the data on this project.
I will be able to hand over the data once we have the signed documents from all team members.

Please sign the document (digitally) and return to myself at kersti.mikkov@rik.ee.

Thank you!

Thank you for choosing this project! ⭐

We are so excited to see what is possible with the data that we have.
If we can use this project to improve the way business is run in Estonia - that would be amazing!

One thing we know for sure is that we will all learn from this project and that this is just the start of all the brilliant things we can do with ML 🙂

Good luck and happy (machine) learning!

The whole e-Business Register team

The data tables

ML_P09_1legal_entities.csv

	Field	Description	Details
General data of the legal entity (as on 13 October 2024)	**le_ps_id**	Pseudonymised id code of a legal entity.	As on 13 October 2024.
	current_status	Current status of the legal entity in the register.	R - registered L - in liquidation N - in bankruptcy K - deleted
	date_registration	Date when the legal entity was first registered in the Estonian business register.	
	date_deletion	Date when the legal entity was deleted from the register.	NA - the entity has not been deleted.
	entity_age	Legal entity's age in years.	As on 13 October 2024 for active entities. For deleted entities - age at the time of deletion.
	legal_form_code	Code for the legal form of the legal entity.	As on 13 October 2024. These two fields are correlated and give the same information.
	legal_form	Legal from in words (in English).	
Main activity	**emtak**	Code representing the main activity of the legal entity (numeric, but should be used as character).	Representing the 5th tier and highest level of accuracy of the main activity of the entity. Legal entities can also have secondary etc. activities - these are not given in this data set. As on 13 October 2024. These two fields are correlated and give the same information.
	activity	Main activity of legal entity in words (in English).	
	activity_field_code	Code for the main activity field of the legal entity (using letters of alphabet).	Representing the 1st tier and lowest level of accuracy of the main activity of the entity. This is the 'mother' of the activity in the previous two rows and is always correlated. There are ~20 fields of activity all together. As on 13 October 2024. These two fields are correlated and give the same information.
	activity_field	Main activity field of the legal entity (in English).	

Annual fiscal year report	**MAA_xxxx_status**	Status of the annual fiscal report of the entity. 1 - the report has been submitted 0 - the report has not been submitted NA - the report is not required	Most legal entities are required by law to submit an annual fiscal year report to the Estonian business register. The report is due 6 months after the end of the financial year of the legal entity. For most legal entities (98%) the fiscal year is the same as a calendar year, ending on 31 December, and so their report due date is on 30 June the following year. The data is given for fiscal years 2019 - 2023.
	MAA_xxxx_late_days	If the report was submitted past the deadline - how many days late was it submitted. 0 - the report was submitted before the deadline 24 - the report was submitted 24 days late NA - the report is not required (or was not late?)	
Warnings	**X2 - xxxx**	Deletion caution: annual report not submitted.	The data has been given for warnings issued in the calendar years 2023 and 2024. Any warnings related to the fiscal year report deficiencies are given with the **fiscal year number** they relate to (as these can be issued several years later). Other warning types are given with the year number of the issue date.
	Y2 - xxxx	A ruling of warning for compulsory dissolution: insufficient net assets	
	Y3 - xxxx	A ruling of warning for compulsory dissolution: insufficient net assets	
	K2 - xxxx	Deletion caution: non-compliance of the contact person	
	HM - xxxx	Warning of a fine.	
	I1 - xxxx	Compulsory dissolution caution: subsection 40(2) of the General Part of the Civil Code Act and section 58 of Commercial Register Act	
	I2 - xxxx	A ruling of warning for compulsory dissolution: the term of the board has terminated	
	GY - xxxx	Compulsory dissolution caution: the legal person does not comply with the requirements established for the legal person by law	
	KI - xxxx	Compulsory dissolution caution	
	G1 - xxxx	Compulsory dissolution caution: deficiencies in the articles of association	
Fines	**H - xxxx**	Penalty payment: - annual report not submitted	The data has been given for fines issued in the calendar years 2023 and 2024. Any fines related to the fiscal year report deficiencies are given with the **fiscal year number** they relate to (as these can be issued several years later). Other fine types are given with the year number of the issue date. The fines listed here are all made out to the legal entity and legal entity is responsible for paying the fine. Any fines issued to the representatives of the company (physical persons) are not given in this data set. NA - no fine has been issued 1 - a fine has been issued and the fine has been paid 0 - a fine has been issued, but the legal entity has not paid the fine
	H2 - xxxx	Repeated penalty payment: annual report not submitted	
	T - xxxx	Penalty payment	
	R - xxxx	Penalty payment: wrong address	
	KO - xxxx	Penalty payment - non-compliance of the contact person	
	KO2 - xxxx	Repeated penalty payment - non-compliance of the contact person	
Tax information	**xxxx_y_sales_tax_reg**	Data on whether the legal entity was registered to pay sales tax during the quarter.	xxxx - year y - quarter of the year jah - was registered for sales tax ei - was not registered for sales tax NA - no information on this legal entity in the taxes data set
	xxxx_y_location	Main location of the legal entity during the quarter.	
	xxxx_y_state_taxes	State taxes payable for the quarter.	In Euros.
	xxxx_y_employment_taxes	Employment taxes payable for the quarter.	In Euros.

	xxxx_y_revenue	Revenue for the quarter.	In Euros.
	xxxx_y_employees	Number of employees.	

ML_P09_phys_persons.csv

Field	Description	Details
pp_ps_id	Pseudonymised id code of a physical person.	
country_of_origin	Country of the id code if the id code is not an Estonian id code.	Empty when Estonian code.
gender	Persons gender based on the Estonian id code.	male / female / unknown
age	Person's age.	Given in years.

ML_P09_3relations.csv

Field	Description	Details
pp_ps_id	Pseudonymised id code of a physical person.	Matches to the 'pp_ps_id' field in the ML_P09_2phys_person table.
le_ps_id	Pseudonymised id code of a legal entity.	Matches to the 'id' field in the ML_P09_1legal_entities table.
starting_date	Starting date of the relation between the person and legal entity.	
ending_date	Ending date of the relation between the person and legal entity.	
role_in_entity	Role the person plays in the legal entity.	

Contacts

For any questions on the data or on how the business register operates, please do get in touch - I am very happy to help!

Best way would be to contact me directly through ML Slack or email me on my work address: kersti.mikkov@rik.ee.

I'm also at Delta most Mondays and Tuesdays, so if you would prefer to catch up in person, I'm sure we can arrange that too 🙂