Bundesministerium
Finanzen

Predictive Analytics Competence Center
bmf.gv.at

Autoencoders

Anomaly Scores as Features in
Supervised Models for Different Tax Areas

Konstantin Posch

(c) http://www.garagezero.eu/20170805193813_gartner-quadrant-big-data/

Roadmap

1. Motivation

2. Autoencoder Basics

3. Applications

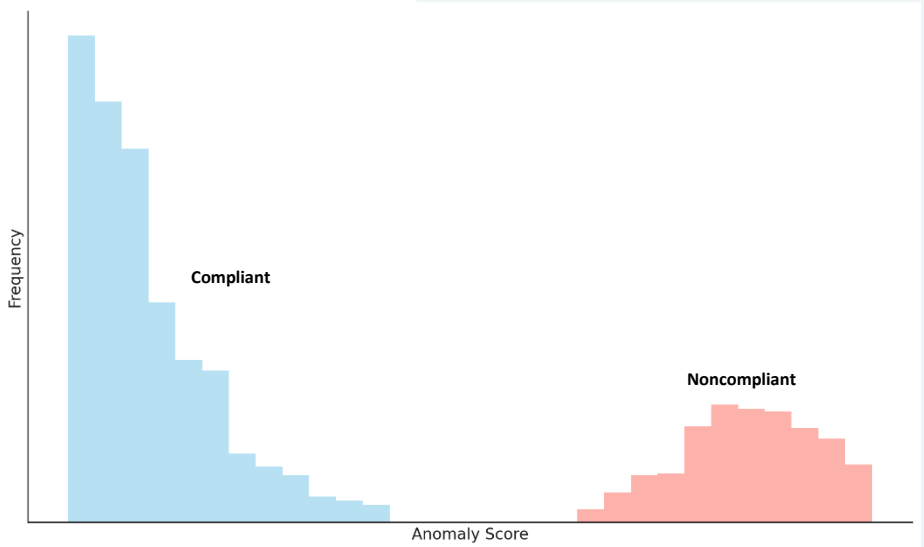


Illustration: Anomaly Score vs. Compliance

1. Motivation

Anomaly Detection for Prevention of Noncompliance

- Anomalous behavior often relates to noncompliant activities
- Anomaly detection is state-of-the-art in fight against credit card fraud
  - PayPal, AMEX, Visa, etc. are known to use anomaly scores to identify fraudulent transactions
- Anomaly scores are common features for risk analysis in diverse tax areas

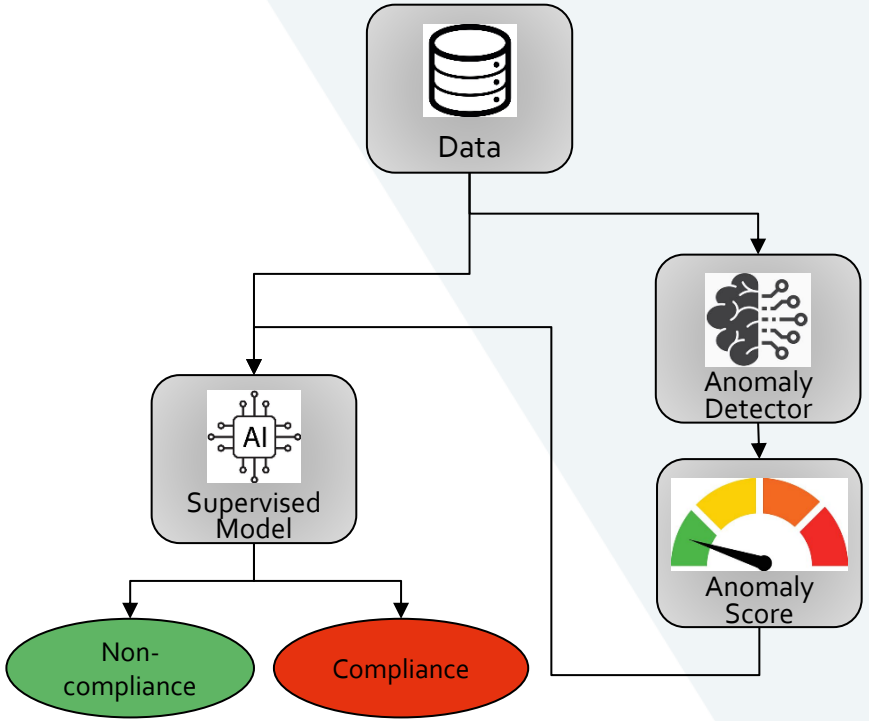
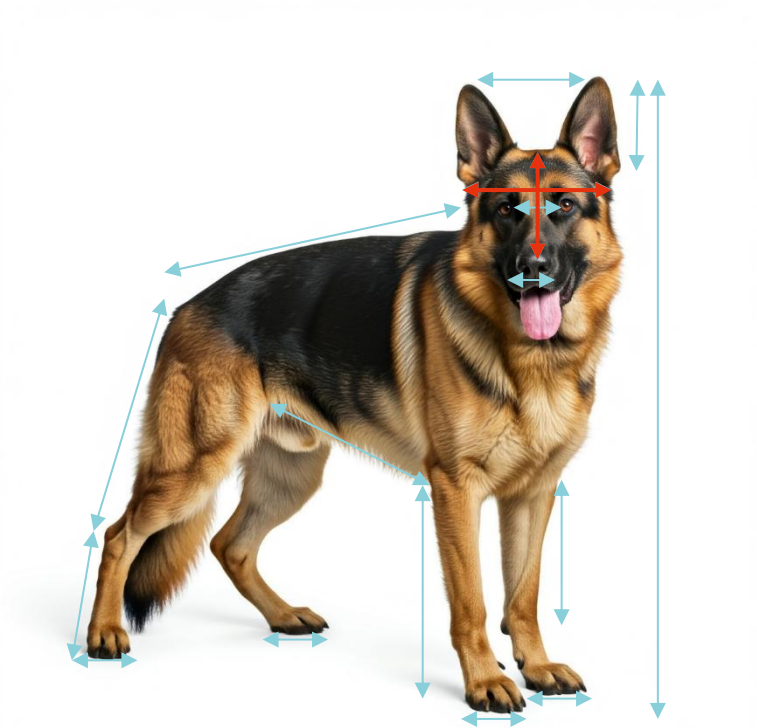


Illustration: Anomaly Score as Feature

2. Autoencoder Basics

Autoencoder – Intuition



Source: Gemini / Imagen 3

Autoencoder – Architecture

- Neural network composed of two components
  - Encoder: Compresses the input data in a low-dimensional representation
  - Decoder: Reconstructs the input data from the low-dimensional representation

- The most important patterns and features are automatically extracted and stored in the compressed representation

- Anomalous data is reconstructed worse than normal data→ Reconstruction error serves as anomaly score

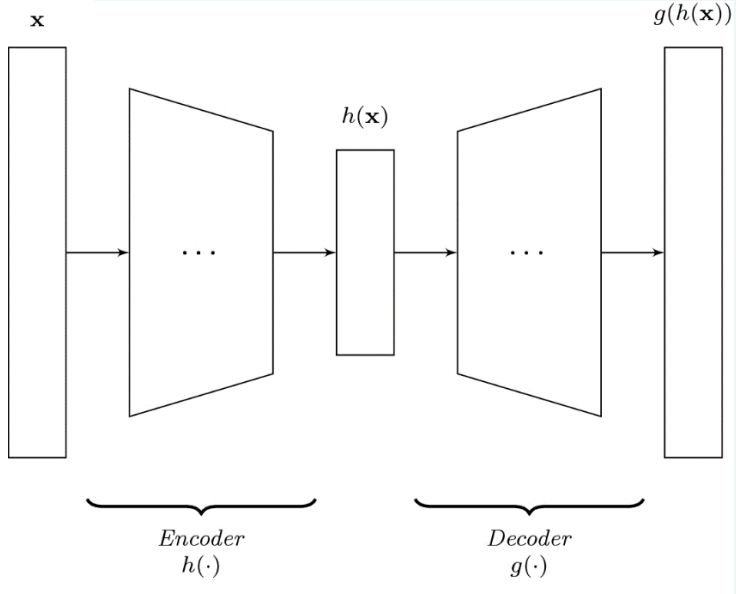


Illustration Autoencoder. Source: *van Engelen, Jesper E. and Holger H. Hoos. "A survey on semi-supervised learning." Machine Learning 109 (2019): 373 - 440*

Autoencoder – Customs and Tax Administration

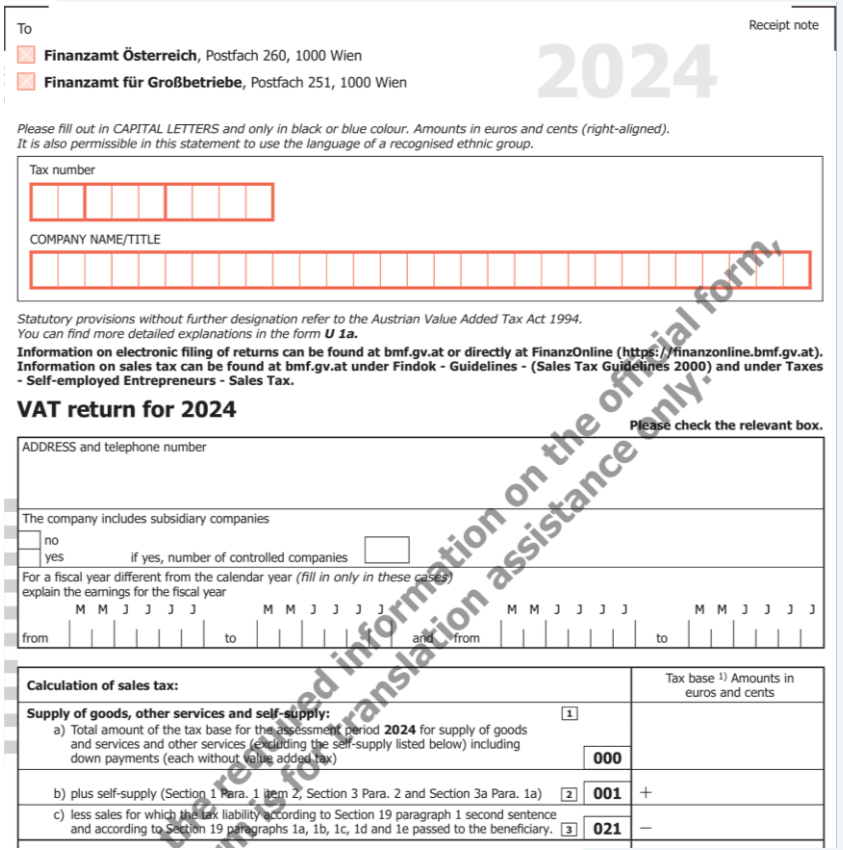
- Autoencoders help to automatically identify unusual tax returns (corporate tax, income tax, VAT, etc.) and customs declarations

1. **Compression of declarations** $\boldsymbol{x} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$ to few key numbers $\boldsymbol{z} = (z_1, \dots, z_m)$
2. **Decompression** of $\boldsymbol{x}$ to $\widehat{\boldsymbol{x}} = (\widehat{\boldsymbol{x}_1}, \dots, \widehat{\boldsymbol{x}_n})$
3. Computation of **reconstruction error**
   - Overall: $\mathrm{Re}Err(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$
   - On feature-level: $\mathrm{Re}Err(x_i, \hat{x}_i) = x_i - \hat{x}_i$

3. Applications

Value Added Tax

- In Austria, companies submit their preliminary VAT return on a monthly/quarterly basis
- It‘s an overview of the value added tax
  - collected from customers (output tax)
  - paid on business-related acquisitions (input tax)
- The difference of the overall collected and paid VAT is the amount of money that has to be paid to or refunded by the tax office



To

Receipt note

Finanzamt Österreich, Postfach 260, 1000 Wien

Finanzamt für Großbetriebe, Postfach 251, 1000 Wien

2024

Please fill out in CAPITAL LETTERS and only in black or blue colour. Amounts in euros and cents (right-aligned).
It is also permissible in this statement to use the language of a recognised ethnic group.

Tax number

COMPANY NAME/TITLE

Statutory provisions without further designation refer to the Austrian Value Added Tax Act 1994.
*You can find more detailed explanations in the form **U 1a.***

Information on electronic filing of returns can be found at bmf.gv.at or directly at FinanzOnline (https://finanzonline.bmf.gv.at).
Information on sales tax can be found at bmf.gv.at under Findok - Guidelines - (Sales Tax Guidelines 2000) and under Taxes - Self-employed Entrepreneurs - Sales Tax.

VAT return for 2024

Please check the relevant box.

ADDRESS and telephone number

The company includes subsidiary companies

no

yes if yes, number of controlled companies

For a fiscal year different from the calendar year *(fill in only in these cases)* explain the earnings for the fiscal year

M M J J J J from M M J J J J to and M M J J J J from M M J J J J to

Calculation of sales tax:		Tax base 1) Amounts in euros and cents
Supply of goods, other services and self-supply: 1 a) Total amount of the tax base for the assessment period **2024** for supply of goods and services and other services (excluding the self-supply listed below) including down payments (each without value added tax)	**000**	
b) plus self-supply (Section 1 Para. 1 item 2, Section 3 Para. 2 and Section 3a Para. 1a) 2	**001**	+
c) less sales for which the tax liability according to Section 19 paragraph 1 second sentence and according to Section 19 paragraphs 1a, 1b, 1c, 1d and 1e passed to the beneficiary. 3	**021**	–

Illustration VAT Return

Value Added Tax – Supervised Models

Currently, we use autoencoder reconstruction errors of VAT data in two models:

- **Risk Scoring: Preliminary VAT Return**
  - Goal: Identification of VAT returns where an audit leads to significant additional tax revenue

- **Risk Scoring: Missing Trader Intra Community (MTIC) Fraud**
  - Goal: Identification of companies involved in MTIC fraud
  - Fraud pattern: Company B (missing trader) buys goods free of input tax from intra-community company A, sells them to a company in the same country and disappears before delivering the collected output tax to the tax office.

Value Added Tax – Data Selection/Preparation

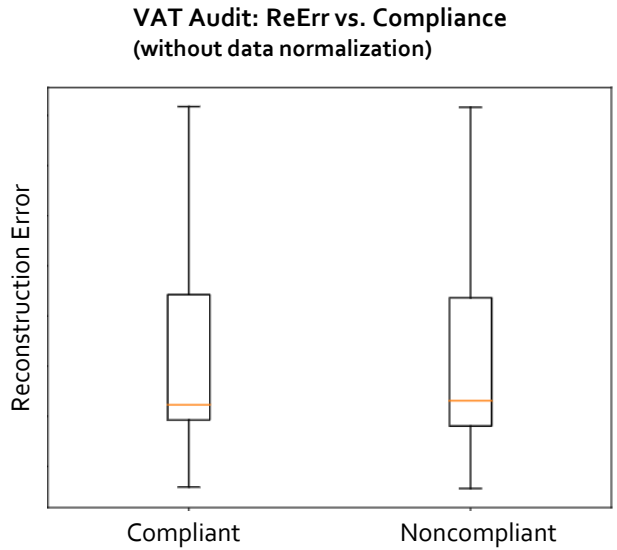
- Selection of actual data
  - Preliminary VAT returns of the previous two years
  - No COVID effects
- Removal of sparse tax numbers (different from zero in less than, e.g., 4% of the cases)
  - Often difficult to reconstruct and of limited utility
- (Optional) Removal of noncompliant VAT returns
  - Some VAT returns are known to be noncompliant from previous tax audits
  - Ideally, the learning should be from normal, i.e., compliant VAT returns

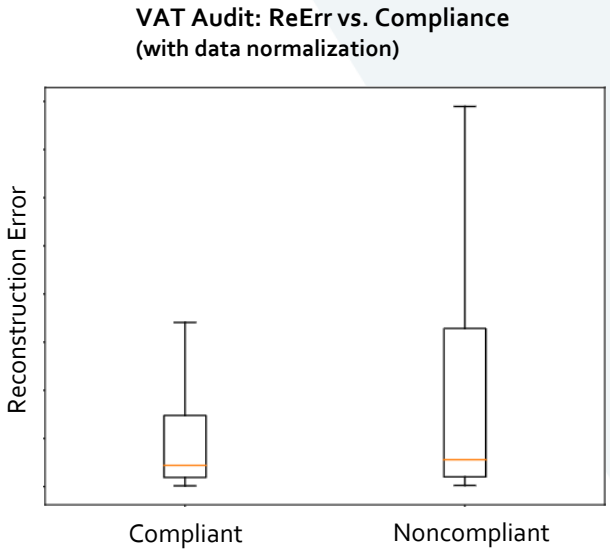
VAT_99	Payment debit/credit
VAT_29	Sales - 10% tax rate
VAT_11	Sales - export supplies
VAT_17	Sales - intra-community supplies
VAT_22	Sales - 20% tax rate
VAT_60	Total amount of input taxes
VAT_65	Input taxes from intra-community acquisition of goods
VAT_72	Intra-community acquisitions - 20% tax rate
VAT_73	Intra-community acquisitions - 10% tax rate

Excerpt of VAT Numbers

Value Added Tax – Data Normalization

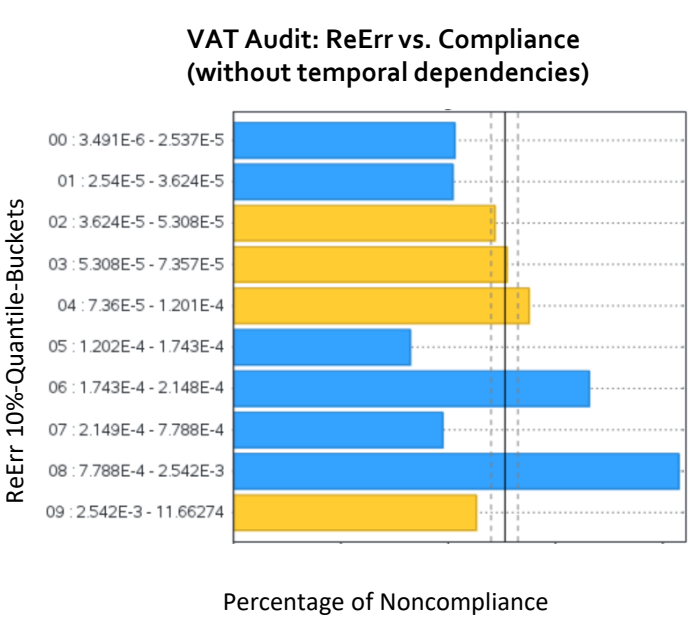
- Key challenge: detection of relevant anomalies and not just any anomalies
- Normalization: divide by sum of all acquisitions and sales at row level to achieve anomaly detection independent of company size

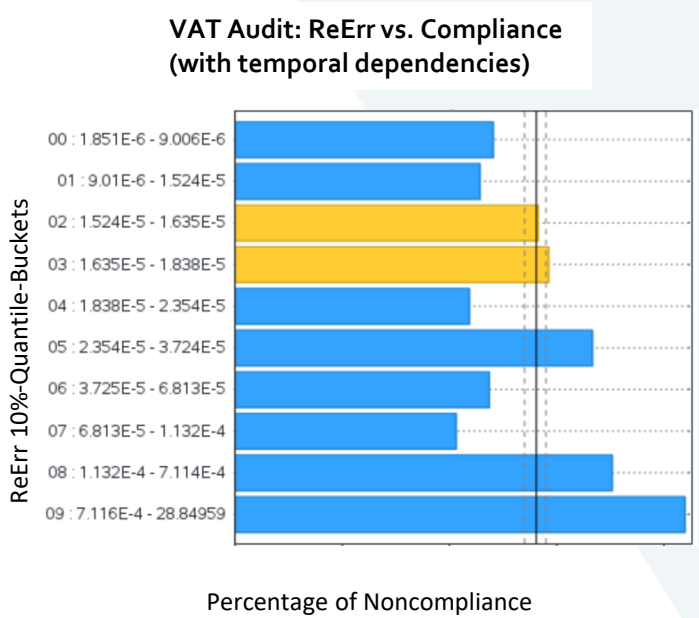




Value Added Tax – Temporal Dependencies

- Idea: Consider the history of tax numbers. Is the current VAT return suspicious compared to past ones?
- Solution: For each tax number, include also its mean value over the previous 12 months



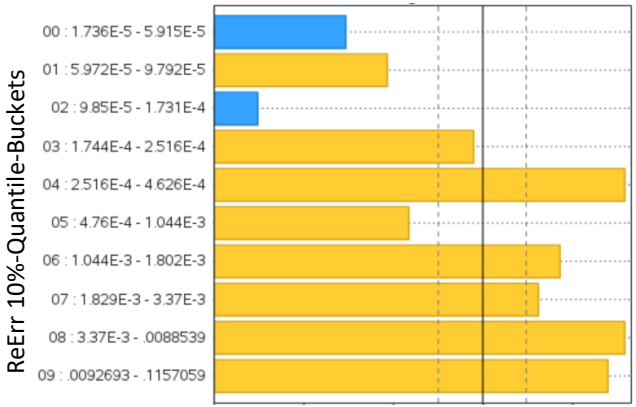


Value Added Tax – Variable Selection

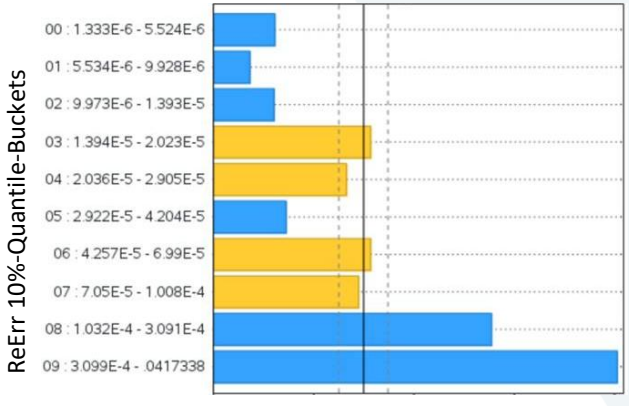
Selection of variables where anomalies correlate with compliance

- Based on expert knowledge
- Based on dependence analysis and retraining

MTIC Fraud: ReErr vs. Fraud
(without variable selection)



MTIC Fraud: ReErr vs. Fraud
(with variable selection)



Value Added Tax – Results

- **Risk Scoring: VAT Return**
  - Feature importance: Reconstruction errors belong to the top 20%
  - Most important ReErrs: VAT_60 (input taxes), VAT_72 (intra-community acquisitions), VAT_22 (sales - 20% tax rate), ...
  - Individual ReErrs exceed the overall one in importance

- **Risk Scoring: Missing Trader Intra Community (MTIC) Fraud**
  - Feature importance: Reconstruction errors belong to the top 25%
  - Most important ReErrs: VAT_65 (input taxes from intra-community acquisitions), VAT_22 (sales - 20% tax rate), ...
  - Evaluations have shown that including ReErrs leads to a significant increase of model accuracy

Import Customs

- Common noncompliant behaviors to reduce customs duties:
  - **Misclassification:** Selection of a commodity code with a lower tariff rate
  - **Undervaluation:** Declaration of a lower customs value (assessment basis for customs duty including invoice amount of the item + transport costs to EU border etc.)

- We perform a risk scoring of items for misclassification and undervaluation

- Autoencoder anomaly scores based on various variables:
  - Commodity code
  - Country of origin
  - Type of transport to EU border (maritime traffic, rail transport, …)
  - Mass (in kg)
  - Customs value
  - Additional import duties (e.g.: anti-dumping duty)
  - …

Import Customs – Data Preparation

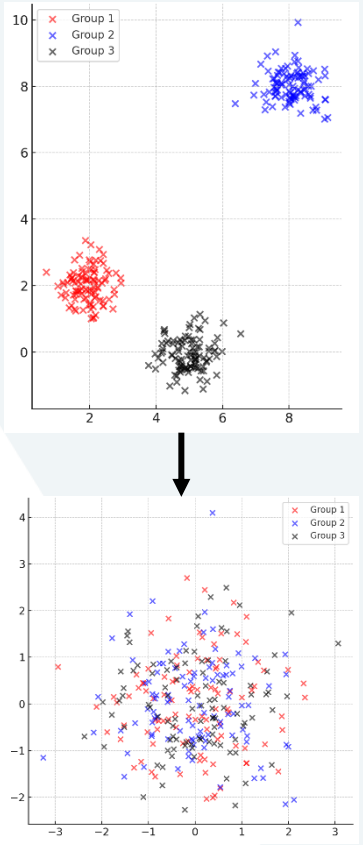
Similar to the VAT data, but additionally:

- Dummy coding of categorical data (merge rare categories into one)
- Standardization with respect to commodity code
  - Commodity code is an essential variable since it determines customs duties
  - Dummy coding is not an option since the number of different categories is huge and imbalanced, which makes it hard to learn useful embeddings
  - Current solution: Smooth standardization of variables per commodity code

Import Customs – Smooth Standardization

- Smooth standardization: $\widehat{\mu_i}^s = \frac{smooth \cdot \hat{\mu} + n_i \cdot \widehat{\mu_i}}{smooth + n_i}$, $\widehat{\sigma_i}^s = \frac{smooth \cdot \hat{\sigma} + n_i \cdot \widehat{\sigma_i}}{smooth + n_i}$
  - $\hat{\mu}$, $\widehat{\mu_i}$: overall and groupwise mean
  - $\hat{\sigma}$, $\widehat{\sigma_i}$: overall and groupwise standard deviation
  - $smooth, n_i$: smoothing parameter and groupwise sample size

- Advantages
  - The influence of the commodity code on size and scale is removed
  - Product groups with extraordinary large/small values are not necessarily considered as anomalous
  - Rare groups get useful mean and variance estimates due to smoothing



(Smooth) Groupwise Standardization

Import Customs – Results

- Feature importance
  - Reconstruction errors belong to the top 25% for undervaluation
  - No notable importance for misclassification

- Most important ReErrs: Transport costs to EU border, type of transport to EU border (rail transport), overall ReErr, etc.

- Next Step: Find a better way to include commodity codes

Corporate Tax & Income Tax

- We perform a risk scoring of
  - corporate tax returns (legal entities, i.e., companies) and
  - income tax returns *(natural persons, e.g., employees, self-employed individuals, etc.)*

  including:
    - Attachments for business income
    - Attachments for renting and leasing of real estate
    - …

- Autoencoders are used for anomaly detection

3. Property withdrawal at book value *(tick where applicable and complete item 8)*

☐ In the fiscal year, one or more properties were removed from the business assets at book value.

4. Profit determination 17

As a rule, income/operating income and expenses/operating expenses ***unsigned*** *must be stated. Only if a ratio results in a negative value, a negative sign ("-") must be indicated.*

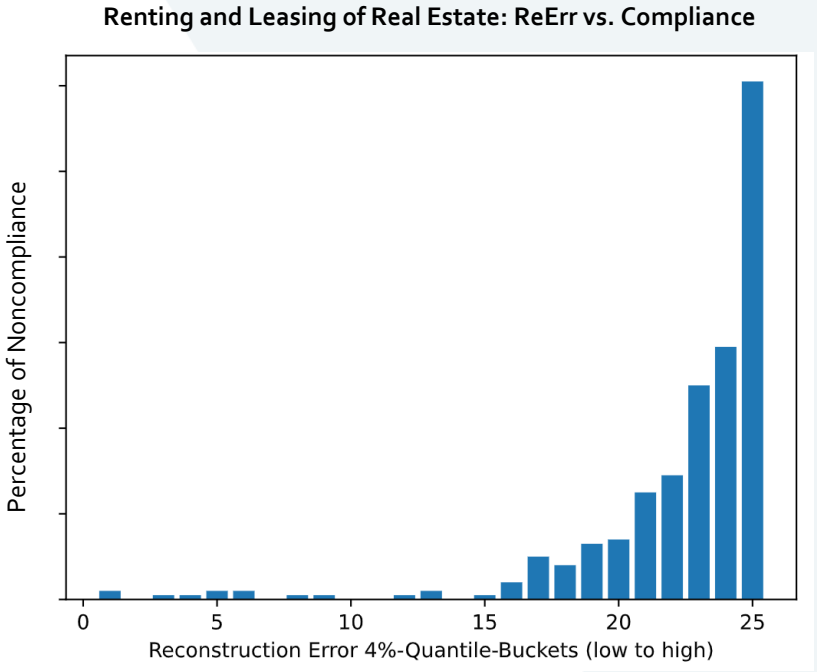
Earnings/income			Amounts in euros and cents
Income/operating income (proceeds from goods/services) excluding those recorded in a notification pursuant to section 109a - EKR 40-44 - including own consumption (withdrawal values from current assets) ***Be aware that:*** *This code must be filled out (section 61 para. 5 FFC). If necessary, enter the value "0".*	18	**9040**	
Earnings/income recorded in a notification pursuant to section 109a EKR 40-44 ***Please note:*** *This code must be filled out (section 61 para. 5 FFC). If necessary, enter the value "0".*	19	**9050**	
Investment income/withdrawal values from fixed assets EKR 460-462 before any resolution to 463-465 or 783	20	**9060**	
Only for balance sheet accountants: Internally produced and capitalised assets EKR 458-459	21	**9070**	
Only for balance sheet accountants: Inventory changes EKR 450-457	22	**9080**	
Other income/operating income (e.g. financial income, profit shares from a silent partnership) - Balance *(For VAT gross system: incl. VAT credit, but without code* ***9093****)*	23	**9090**	
Only for VAT gross system: VAT paid for supplies and other services *(Attention: Only fill in if the operating income is stated without VAT)*	24	**9093**	
Total expenses/operating expenses *(does not have to be filled in)*			
Expenses/Operating Expenses			
Goods, raw materials, auxiliary materials EKR 500-539, 580	25	**9100**	
Provided personnel (external personnel) and external services EKR 570-579, 581, 750-753	26	**9110**	
Personnel expenses ("own personnel") EKR 60-68	27	**9120**	
Depreciation of fixed assets (e.g. depreciation, low-value assets, EKR 700 - 708), unless they are under code **9134** and/or **9135**, must be recorded.	28	**9130**	
Declining depreciation for wear (section 7 para. 1a)	29	**9134**	
Accelerated depreciation of buildings (section 8 para. 1a and section 124 para 451)	30	**9135**	
Only for balance sheet accountants: Depreciation of current assets to the extent that they exceed the customary depreciations in the company - EKR 709 - and allowances for receivables, to the extent that they should not be recorded in code **9142**	31	**9140**	
Allocation/reversal of flat-rate value adjustment on receivables ***Be aware that****: For reversals, the amount must be entered with a negative sign.*	32	**9142**	
Maintenance (maintenance costs) for buildings EKR 72	33	**9150**	
Travel expenses including mileage allowance and daily allowances (but not actua motor vehicle costs) EKR 734-737	34	**9160**	
Flat rate of 50% of the costs one week, month or annual mass transit pass	35	**9165**	
Actual motor vehicle costs (without depreciation for wear and tear, leasing and mileage allowance) EKR 733-737	36	**9170**	
Rental and leasing expenses EKR 740-743, 746-747	37	**9180**	
Commissions to third parties, licence fees EKR 754-757, 748-749	38	**9190**	
Advertising and representation expenses, donations, tips not to be recorded under codes **9243** to **9209** EKR 765-769	39	**9200**	
Book value of disposed assets EKR 782	40	**9210**	
Work room *No entry may be made under code* ***9215****,* ***9216*** *or* ***9217****. Can only be deducted if the study is the centre of all business activity.*	41	**9275**	

Excerpt of Income Tax Return

Corporate Tax & Income Tax - Results

- Overall reconstruction error belongs to the most important variables in supervised models

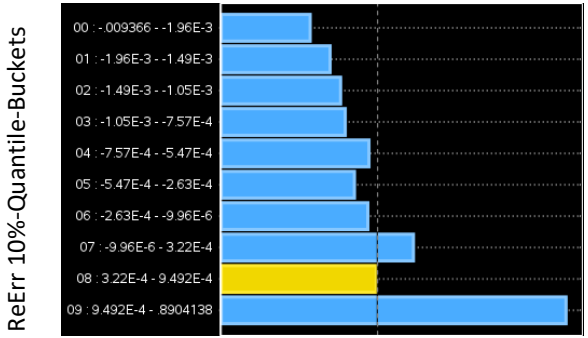
- Strong correlation with the target variable as illustrated in the figure on the right



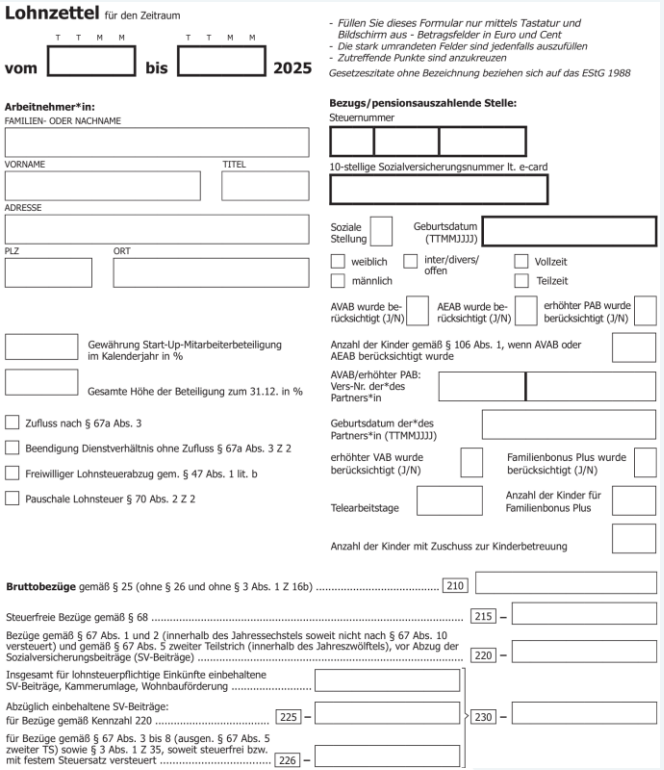
Pay Slips

- We perform a risk scoring of pay slips considering
  - wage-related taxes and
  - social security insurance contributions
- Autoencoders are used for anomaly detection
- Reconstruction errors belong to the most important variables in supervised models

Pay Slip – Tax-Free Income : ReErr vs. Compliance



Percentage of Noncompliance



Lohnzettel für den Zeitraum

vom T T M M bis T T M M 2025

- *Füllen Sie dieses Formular nur mittels Tastatur und Bildschirm aus - Betragsfelder in Euro und Cent*
- *Die stark umrandeten Felder sind jedenfalls auszufüllen*
- *Zutreffende Punkte sind anzukreuzen*

Gesetzeszitate ohne Bezeichnung beziehen sich auf das EStG 1988

Arbeitnehmer*in:
FAMILIEN- ODER NACHNAME
VORNAME TITEL
ADRESSE
PLZ ORT

Bezugs/pensionsauszahlende Stelle:
Steuernummer
10-stellige Sozialversicherungsnummer lt. e-card
Soziale Stellung Geburtsdatum (TTMMJJJJ)
weiblich inter/divers/offen Vollzeit
männlich Teilzeit
AVAB wurde berücksichtigt (J/N) AEAB wurde berücksichtigt (J/N) erhöhter PAB wurde berücksichtigt (J/N)

Gewährung Start-Up-Mitarbeiterbeteiligung im Kalenderjahr in %
Gesamte Höhe der Beteiligung zum 31.12. in %
Zufluss nach § 67a Abs. 3
Beendigung Dienstverhältnis ohne Zufluss § 67a Abs. 3 Z 2
Freiwilliger Lohnsteuerabzug gem. § 47 Abs. 1 lit. b
Pauschale Lohnsteuer § 70 Abs. 2 Z 2

Anzahl der Kinder gemäß § 106 Abs. 1, wenn AVAB oder AEAB berücksichtigt wurde
AVAB/erhöhter PAB: Vers-Nr. der*des Partners*in
Geburtsdatum der*des Partners*in (TTMMJJJJ)
erhöhter VAB wurde berücksichtigt (J/N)
Familienbonus Plus wurde berücksichtigt (J/N)
Telearbeitstage
Anzahl der Kinder für Familienbonus Plus
Anzahl der Kinder mit Zuschuss zur Kinderbetreuung

Bruttobezüge gemäß § 25 (ohne § 26 und ohne § 3 Abs. 1 Z 16b) 210
Steuerfreie Bezüge gemäß § 68 ..... 215 –
Bezüge gemäß § 67 Abs. 1 und 2 (innerhalb des Jahressechstels soweit nicht nach § 67 Abs. 10 versteuert) und gemäß § 67 Abs. 5 zweiter Teilstrich (innerhalb des Jahreszwölftels), vor Abzug der Sozialversicherungsbeiträge (SV-Beiträge) ..... 220 –
Insgesamt für lohnsteuerpflichtige Einkünfte einbehaltene SV-Beiträge, Kammerumlage, Wohnbauförderung .....
Abzüglich einbehaltene SV-Beiträge: für Bezüge gemäß Kennzahl 220 ..... 225 –
für Bezüge gemäß § 67 Abs. 3 bis 8 (ausgen. § 67 Abs. 5 zweiter TS) sowie § 3 Abs. 1 Z 35, soweit steuerfrei bzw. mit festem Steuersatz versteuert ..... 226 –
230 –

Excerpt – Pay Slip

Final Notes

- Autoencoder architectures
  - We used rather small architectures
  - Between 2 and 7 hidden layers
  - Less than 100 nodes with tanh activation per hidden layer

- Challenges to look for alternative solutions
  - Sparse variables – Current solution: removal with sparsity threshold
  - Categorical data with many classes – Current solution: smooth standardization

- To try out
  - Consideration of full temporal history of variables
  - First tries without models specifically designed for time series data did not show advantages

Thank you for your attention!