Machine Learning Using Tax Information

Improving Entity Classification on IRS Form 1041

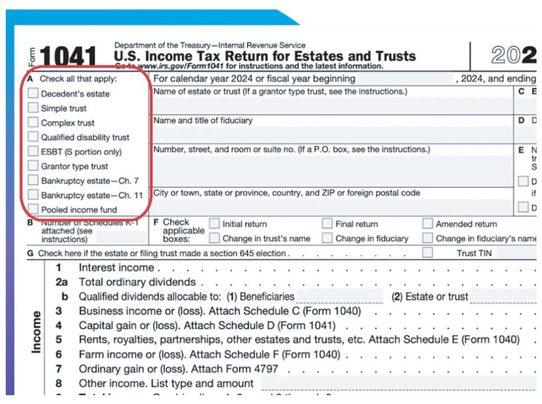
IRS Statistics of Income

Introduction

The Form 1041, U.S. Income Tax Return for Estates and Trusts is used by the fiduciary of an entity (trust, domestic decedent's estate, or bankruptcy estate) to report the income, deductions, gains, losses, and tax liability of that entity.

Types of entities that file this form:

- Simple Trust
- Complex Trust
- Qualified Disability Trust
- Decedent's Estate
- Taxable Grantor Trusts
- Non-Taxable Grantor Trusts
- Bankruptcy Estates (Ch.7, Ch.11)
- Pooled Income Funds
- (ESBT(S) portion)



One or more boxes can be checked in Box A.

Entity type determines tax liability, exemption amounts, filing requirements, and how income is distributed and taxed to beneficiaries.

Entities have complex structures:

- Strict reporting requirements and tax obligations.
- Laws governing them allow for specifications in the trust instruments that may allow certain deductions or exemptions to be claimed or allow changes in reporting requirements for a given year that are not typical.

Examples:

- A Complex Trust is entitled to an exemption of $100 but may be allowed $300 in exemption if the trust instrument requires all income to be distributed currently, even if the full distribution to the beneficiaries does not occur during that tax year.
- Grantor trusts may include specific provisions regarding the timing of distributions made to beneficiaries which can impact the tax reporting requirements for the tax year.

Background & Problem Statement

The SOI Annual 1041 study uses tax return data extracted from the Compliance Data Warehouse (CDW) and processed end-to end through computational workflows. Since the study does not involve manual editing, Box A (Entity type selected) information is not directly available.

Current Entity Assignment

- Records contain coded fields
- A rule-based program assigns entity labels from those codes
- Labels define the existing assignment system:
  - Incomplete across the dataset
  - Vary in structural clarity

Approximately 14–15% of returns in a calendar year are missing codes and an additional 6–10% cannot reliably be assigned labels using available codes—leading to about 20–25% of returns in a calendar year lacking complete, reliable entity information.

Approach:

Can we use machine learning (ML) to learn patterns in labeled, non-text form entries that define entities and apply them to new records?

Constraint: Labeling System structure

- Some labels correspond to a single entity
- Some labels combine multiple entities into a single category
- Supervised learning requires clearly defined target classes.

Objectives

Can the model learn?

Assess model performance under well-defined categories

What does separability depend on?

Compare separability under atomic and compositional label structures

Is the model applicable?

Apply the trained model to unlabeled data and examine model behavior

METHODOLOGY

Data and Label Structure

- Population of Form 1041 returns from Tax Year 2022.
- 8 entity types defined in current system.
- 4 categories selected for supervised modeling (most structurally well-defined):
  - Decedent Estate
  - Qualified Disability Trust
  - Taxable Grantor Trust
  - Non-Taxable Grantor Trust
  - Remaining categories involve label ambiguity

Modeling & Evaluation Strategy

- Stratified train/test split
- Random undersampling for balanced training
- Balanced test evaluation to assess class-level separability and structured confusion
- Models compared: Multinomial Logistic Regression, Random Forest (RF), Support Vector Machine (SVM), XGBoost.
- Accuracy and F1 reported

Results: Learnability

Performance Across Supervised Models

Model :[DE, QD TG, NTG]

Method	Accuracy (%)	F1 score
Random Forest	94.7	0.95
Support Vector Machine	78.3	0.77
Multinomial Logistic Regression	92.1	0.92
XGBoost	95.1	0.95

DE: Decedent Estate
QD: Qualified Disability Trust
TG: Taxable Grantor Trust
NTG: Non-Taxable Grantor Trust

- Nonlinear tree ensembles outperform Support Vector Machine
- Random Forest and XGBoost achieve comparable high performance
- Random Forest selected for:
  - Comparable accuracy
  - Greater robustness and interpretability
  - Clear global variable importance

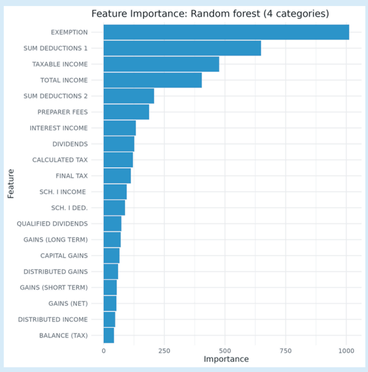
Core four-category performance

Model :[DE, QD TG, NTG]

	Actual Class			
Predicted	DE	NTG	QD	TG
DE	370	0	4	30
NTG	17	401	2	1
QD	0	0	398	1
TG	29	0	4	398

Accuracy : 94.7%; F1 score: 0.95

- Strong overall performance
- Mild misclassification reflects label ambiguity
- Variable importance aligns with domain expectations



Results: Structural Sensitivity

Atomic vs Compositional Label Comparison

Atomic Label

Model [DE, QD]

	Actual Class		Percent
Predicted class	DE	QD	MisClassified (%)
DE	488	5	1
QD	0	495	0

DE: Decedent Estate
QD: Qualified Disability Trust
TG: Taxable Grantor Trust
NTG: Non-Taxable Grantor Trust
ST: Simple Trust
CT: Complex Trust

- **Accuracy = 99.5%**
- Label:
  DE ↔ DE
  QD ↔ QD

Compositional Label

Model : [ST, CT, TG, NTG]

	Actual Class				Percent
Predicted	CT	NTG	ST	TG	MisClassified (%)
CT	353	0	10	52	15
NTG	22	397	6	0	6.6
ST	22	0	367	38	14
TG	30	1	17	324	13

- **Accuracy = 88.3%**
- Label:
  TG ↔ CT/ST
  NTG ↔ CT/ST

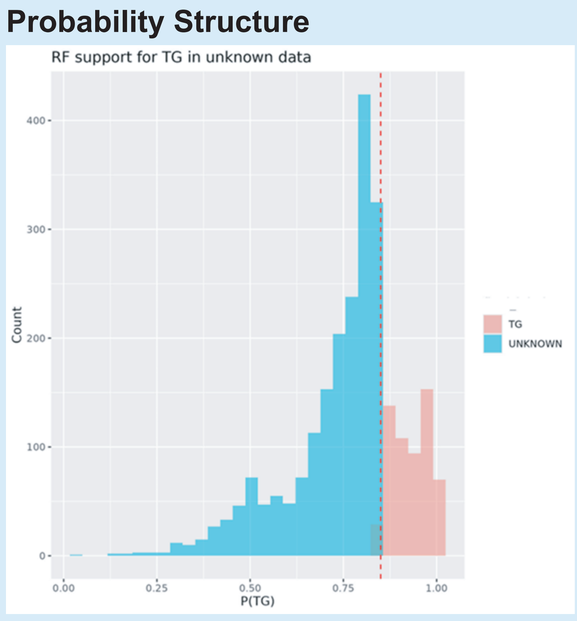
- Confusion patterns reflect label structure and aligns with known label complexity
- Label structure affects separability
- Confusion regions may represent candidates for multi-category labeling or boundary refinement

Results: Projection to Unlabeled Data

Projection Approach

- Applied trained core model to previously unlabeled dataset
- Restricted to records with adequate feature content to evaluate structural applicability of the model
- All four categories were evaluated; after applying minimum threshold to the top predicted class, only Class TG met the criterion.
- Records classified as Class TG if P(TG) ≥ 0.85
- Remaining records labeled UNKNOWN

Probability Structure



Prediction	Count
TG	592
UNKNOWN	1908

- Coherent subset of records strongly align with the learned pattern for class TG
- Enables targeted validation of conservatively identified cases

CONCLUSION

What the study shows

- Random Forest can learn meaningful entity distinctions from Form 1041 data.
- Class separability reflects underlying label structure.
- Learned category patterns are detectable in previously unlabeled data.
- High-confidence candidate cases can be identified conservatively.

Deployment Considerations

- Validation of candidate cases required.
- Threshold calibration under natural prevalence.
- Handling of sparse or ambiguous cases.
- Additional label refinement can expand coverage.

Contact:
Sudeshna Roy
Email: Sudeshna.Roy@irs.gov