Tax Filing Patterns via Classification & Clustering

Jesus Casas, Victoria Miteva, Luis Olmos, Abdul Hasib Safi, and Luis Cedeno
Department of Mathematics, California State University Northridge

Abstract. This project applies machine learning techniques in R to analyze taxpayer behavior using real-world data provided by Nesban Solutions, a local tax and accounting firm. The primary objective is to identify distinct groups of taxpayers based on filing behavior through unsupervised learning (clustering), and to predict refund versus tax owed outcomes using supervised classification models. Key variables include filing date, income, dependents, tax withheld, employment type, and filing status. By clustering taxpayers, we uncover behavioral patterns such as early filers, high-income owners, and refund-maximizing families. Classification models such as logistic regression and random forest are evaluated to predict refund likelihood, with attention to class imbalance. Insights from this analysis can support fraud detection, personalized tax strategy recommendations, and audit prioritization, offering practical value to both clients and tax professionals.

1 Introduction

Taxpayer behavior plays a critical role in shaping both individual financial outcomes and institutional tax strategy. The ability to identify behavioral patterns and predict tax outcomes has significant implications for compliance auditing, fraud detection, and personalized tax planning. This study explores the application of machine learning techniques to analyze taxpayer filing behavior using real-world data obtained from Nesban Solutions, a regional accounting and tax services firm.

The primary objectives of this research are twofold: (1) to segment taxpayers based on their filing and financial characteristics using unsupervised learning techniques, and (2) to develop predictive models that classify whether a taxpayer is likely to receive a refund or owe taxes. The dataset comprises variables such as filing date, income level, employment type, filing status, number of dependents, tax withheld, and total tax liability, features commonly available in tax return records.

Clustering algorithms are employed to uncover latent behavioral groupings among taxpayers, potentially revealing profiles such as early filers, late filers, consistent owers, or refund-maximizing households. Supervised classification models, including logistic regression and random forest, are implemented to predict refund status. Special attention is given to model evaluation under class imbalance, utilizing metrics such as precision-recall AUC and F1 score.

This research contributes to the growing intersection of data science and financial services by demonstrating how machine learning can enhance decision-making in the domain of tax compliance and client segmentation.

2 Methodology

This section outlines the data preprocessing, clustering, and classification techniques employed in the analysis. The overall approach is divided into three phases: data preparation, unsupervised clustering, and supervised classification modeling.

2.1 Data Preprocessing

Raw data collected from Nesban Solutions included both structured numeric and categorical fields. To ensure model readiness, the dataset underwent the following preprocessing steps:

- **Data Cleaning:** Inconsistent formats (e.g., commas in numeric fields, missing or malformed entries) were corrected. Dates were converted to R Date types and parsed appropriately.
- **Feature Engineering:**
  - Filing_Date was transformed into a continuous variable Filing_DayOfYear to encode seasonal filing behavior.
  - A binary classification target variable refund_flag was created based on the sign of the refund/owed amount (1 = refund, 0 = owes).
- **Categorical Encoding**: Employment type and filing status were one-hot encoded using dummyVars() to allow inclusion in machine learning models.
- **Scaling:** All continuous variables were standardized to zero mean and unit variance prior to clustering to ensure balanced feature contribution.

2.2 Exploratory Data Analysis

Prior to modeling, several visualization techniques were used to assess feature distributions, multicollinearity, and the overall clusterability of the dataset.

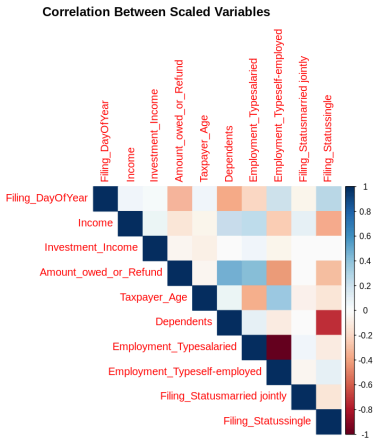

Fig 1. Correlation Heatmap: A correlation matrix was generated to detect collinearity between numeric features such as income, age, investment income, and refund amount.

This correlation matrix helps identify multicollinearity and linear associations between numeric features/ Income shows a strong positive correlation with Amount Owed or Refunded and Tax Withheld, which is expected. Number of Dependents correlates positively with Refund Amount, suggesting a possible link between family size and refund eligibility. Taxpayer Age appears moderately correlated with employment type and filing status.

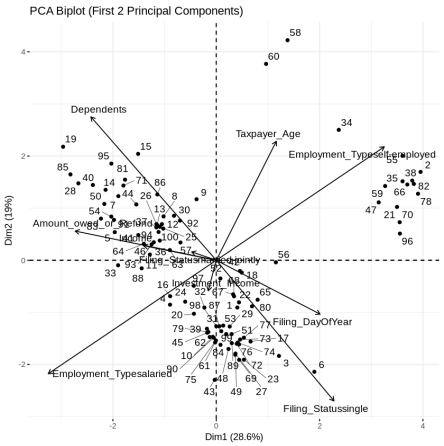

Fig 2. PCA was conducted to reduce dimensionality and visualize data structure in 2D space. The biplot in Figure 2 displays the first two principal components along with directional loadings for each feature.

The PCA biplot displays the first two principal components, which together explain a significant portion of the variance in the dataset (approximately 47.6%). Several trends are immediately

observable:

- Employment Type is a major driver of variance along Dim1, separating salaried and self-employed individuals.
- Taxpayer Age, Dependents, and Filing Day of Year contribute meaningfully along Dim2.

Variables such as Amount Owed or Refunded and Income are moderately aligned with the first component, suggesting their importance in behavioral differentiation.

Taxpayers are not uniformly distributed in the PCA space, suggesting latent groupings suitable for clustering.

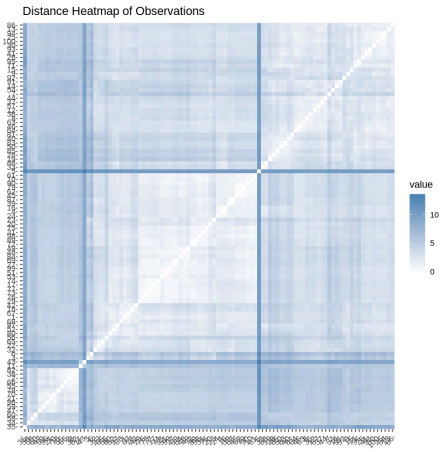

Fig 3. Distance Matrix Visualization: A heatmap of pairwise distances was computed using Euclidean distance, providing an initial view of the density and separation among taxpayers.

Figure 3 shows that brighter areas indicate greater distances, while darker squares highlight clusters of taxpayers who are behaviorally or financially similar. The presence of multiple dark diagonal blocks supports the hypothesis that the data is clusterable. Several sharp breaks in the gradient suggest well-separated groups, which aligns with the DBSCAN and k-means clustering assumptions tested later.

These exploratory tools guided model decisions but were not used to make inferences at this stage.

2.3 Clustering Analysis

The purpose of clustering in this study was to segment taxpayers into distinct behavioral groups based on their financial attributes and filing patterns. Several clustering algorithms were applied and evaluated to identify natural groupings within the data.

2.3.1 K-Means Clustering

The K-means algorithm was applied to the scaled dataset with varying values of k (number of clusters). The optimal k was selected using the elbow method (see Figure 4 below), which plots the total within-cluster sum of squares (WSS) against the number of clusters.

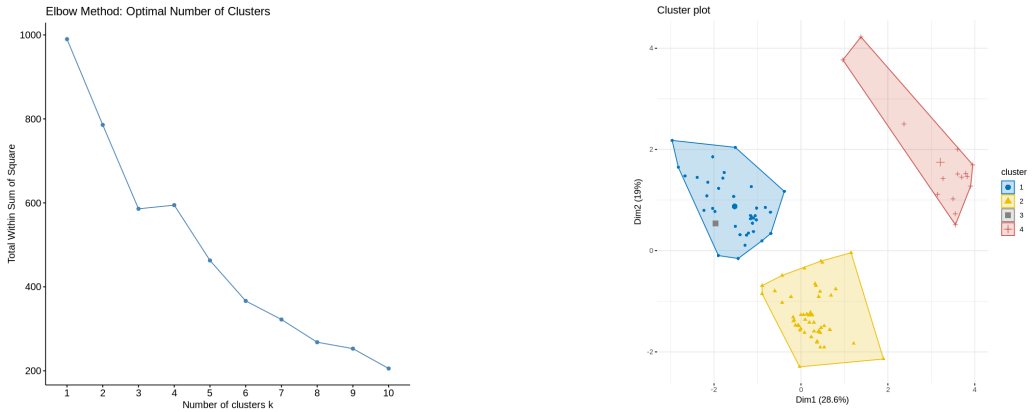

Fig 4. On the left, the "elbow" appears at k = 4, where the marginal gain in WSS reduction begins to plateau. This suggests that four clusters balance between overfitting and under-segmentation. On the right, a PCA-based cluster visualization shows the resulting cluster separation.

Clusters appear well-separated in PCA space, indicating clear behavioral patterns. Each cluster forms a relatively compact group, supporting the assumption of isotropic (spherical) clusters required by K-means.

2.3.2 Hierarchical Clustering

Fig 5. On the left, hierarchical clustering using Ward's method was also performed, generating a dendrogram. This method builds a dendrogram representing nested clusters, which can be cut at any level to form k groups. On the right, the clusters were projected in PCA space revealing three dominant and one small, compact group.

A cut at four groups (matching the K-means result) was used for consistency. This method helped confirm that multiple levels of behavioral similarity exist and that smaller subgroups could be explored further in future work.

2.3.3 DBSCAN (Density-Based Clustering)

To detect clusters of varying shape and handle outliers, DBSCAN was employed. The optimal eps parameter was chosen using the k-distance plot (Figure 6):

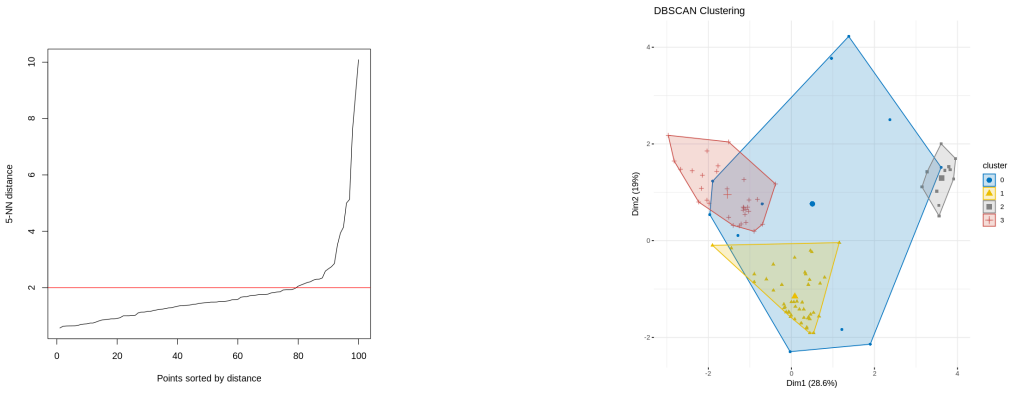

Fig 6. On the left, the sharp bend in the curve occurs around eps = 2, which was chosen as the optimal density threshold. Unlike K-means, DBSCAN uncovered more organic groupings and was able to isolate sparse outlier points, which may represent unusual filing behaviors.

Using eps = 2 and minPts = 5, DBSCAN identified four primary clusters along with noise points labeled as cluster 0. The clustering result is shown in Figure 7.

2.4 Classification Analysis

The second component of this project aimed to predict whether a taxpayer would receive a refund (class 1) or owe money (class 0) based on financial and demographic attributes. This binary classification task was modeled using both linear and non-linear supervised learning algorithms, evaluated with an emphasis on class imbalance.

The binary response variable refund_flag was derived from the Amount_owed_or_Refund column:

- 1 indicates a taxpayer received a refund (positive value)
- 0 indicates a taxpayer owed taxes or broke even

Predictor variables included:

- Numeric: income, investment income, age, dependents, filing date (converted to day-of-year)
- Categorical: employment type and filing status (one-hot encoded)

All features were scaled where appropriate. The dataset was split into 80% training and 20% test sets using createDataPartition() from the caret package.

2.4.1 Handling Class Imbalance

The second component of this project aimed to predict whether a taxpayer would receive a refund (class 1) or owe money (class 0) based on financial and demographic attributes. This binary classification task was modeled using both linear and non-linear supervised learning algorithms, evaluated with an emphasis on class imbalance.

Preliminary analysis revealed class imbalance, with refund cases (class 1) being more common than owed cases (class 0) (35% class 0 vs. 65% class 1). To mitigate performance degradation on the minority class, two balancing strategies were applied:

- **Weighted Logistic Regression:** Weights were inversely proportional to class

frequencies.

- **Upsampling for Random Forest:** The minority class was oversampled using caret::upSample() to balance the training set.

2.4.2 Logistic Regression (GLM)

Two logistic regression models were trained using the glm() function:

(a) Baseline Logistic Regression

- Formula: refund_flag ~ .
- Family: binomial
- No class weighting

(b) Weighted Logistic Regression

- Used class weights based on inverse class frequency:
  - `class_weights <- total / (2 * class_counts)`
- Weights were passed using the weights = argument in glm()
- Same formula and binomial family

This approach emphasized misclassifying the minority class (those who owe) more heavily in the likelihood function.

2.4.3 Random Forest

Random forests were trained using the randomForest package, each with ntree = 100.

(a) Standard Random Forest

- Trained using the unmodified training set
- Model input used the matrix interface:
  - `randomForest(x = x_train, y = y_train, ntree = 100)`

(b) Upsampled Random Forest

- The minority class (refund_flag = 0) was upsampled using:
  - `upSample(x = ..., y = ..., yname = "refund_flag")`
- The upsampled data was then used to train the model using the same randomForest() call

This strategy ensured equal class representation during training, potentially improving the model's sensitivity to class 0 (tax owed).

2.5 Evaluation Metrics

To evaluate model and clustering performance, a combination of internal validation metrics and

classification diagnostics were applied.

Clustering models (K-means, DBSCAN, GMM, hierarchical) were evaluated using:

- Silhouette Score: Measures cluster cohesion and separation.
- Calinski-Harabasz Index: Assesses between- vs. within-cluster dispersion.
- Dunn Index: Evaluates compactness and separation of clusters.

These metrics enabled direct comparison of clustering structure and guided model selection.

Classification models (logistic regression, random forest) were assessed using:

- Confusion Matrix: Provides accuracy, sensitivity, specificity, and precision.
- F1 Score: Balances precision and recall, especially valuable for class imbalance.
- ROC AUC: Measures ranking performance across thresholds.
- PR AUC: Emphasizes precision-recall tradeoff for the minority class (class 0 = owes).

These metrics were computed using predictions on the holdout test set, and were later compared across model variations to select the best-performing approach for the classification task.

3 Data Analysis

This section presents the empirical results of both the unsupervised clustering and supervised classification models applied to the taxpayer dataset. The analysis focuses on evaluating the effectiveness of different algorithms in segmenting taxpayers by filing behavior and predicting refund outcomes, using the evaluation metrics defined in the methodology.

Clustering Method	Silhouette Score	Calinski-Harabasz Index	Dunn Index
K-means	0.337	33.1	0.156
GMM	0.228	16.7	0.043
Hierarchical	0.331	32.6	0.156
DBSCAN	**0.409**	**64.3**	**0.375**

Table 1: The quality of the clustering results.

DBSCAN achieved the highest silhouette score (0.409), indicating that it produced the most well-separated and cohesive clusters. DBSCAN also outperformed other methods with a score of 64.3 in Calinski-Harabasz Index, suggesting a strong between-cluster separation. DBSCAN achieved the highest value (0.375) in Dunn Index, reinforcing its ability to form compact, distinct clusters while effectively isolating outliers. These metrics collectively identify DBSCAN as the most appropriate algorithm for this dataset.

Metric	Logistic (Base)	Random Forest (Base)	Logistic (Weighted)	Random Forest (Upsampled)
Accuracy	0.80	0.85	0.90	**0.95**
F1-score	0.667	0.7273	0.8571	**0.9231**

Table 2: Classification model comparison (Base + Balanced)

While the upsampled random forest achieved the highest accuracy (95%), accuracy alone is insufficient in the presence of class imbalance. The F1 score for class 0, which balances precision and recall, is a more appropriate measure. Both balanced models improved substantially:

- Weighted logistic regression: F1 increased from 0.6667 to 0.8571
- Upsampled random forest: F1 rose from 0.7273 to 0.9231, with perfect precision

This highlights the effectiveness of class balancing in improving minority class detection.

Although random forest had the highest ROC AUC (0.868), it suffered from a low PR AUC (0.239), reflecting poor performance for class 0 at practical thresholds. In contrast, logistic regression achieved a much stronger PR AUC of 0.825, closely aligned with its ROC AUC (0.824), indicating better-calibrated probabilities. Ultimately, while logistic regression was more balanced and interpretable, the upsampled random forest outperformed it across all key metrics, making it the preferred model.

4 Conclusion

This project applied clustering and classification techniques to analyze taxpayer behavior and predict refund outcomes using real data from Nesban Solutions. DBSCAN clustering revealed four distinct taxpayer groups, including refund-maximizing families and low-income filers who owe taxes.

For classification, both logistic regression and random forest were tested with class balancing. While logistic regression provided well-calibrated predictions, the upsampled random forest model achieved the highest overall performance, including 95% accuracy and an F1 score of 0.9231 for identifying taxpayers who owe. These results offer Nesban Solutions actionable insights for *client segmentation*, *audit prioritization*, and *personalized tax strategy*.

Appendix

1. **Dataset**

2. **Colab Notebook**

3. **Slides**

Cluster	Count	Avg Income	Avg Refund	Avg Age	Avg Dependents	Segment Description
1	46	$26,992	+892	29.2	0.22	Young filers with modest refunds
2	11	$12,559	-$1,567	46.4	0.09	Older low-income group who owe
3	32	$36,378	+$3,399	36.2	1.75	Families with dependents and high refunds

Table 3: These profiles highlight distinct behavioral segments, including refund-maximizing families, at-risk filers, and younger individuals receiving modest refunds. The presence of noise points further demonstrates DBSCAN's strength in detecting irregular or anomalous behavior that may be of particular interest to the firm.