Project Design.

Classification and Predictive Analytics

on the Ontario interest rates.

Nara Galliamova

501143606

Ashok Bhowmick, Ceni Babaoglu, Sedef Akinli Kocak,

Uzair Ahmad, Zekiye Erdem

January 26, 2026

Table of Contents

Problem Statement and Context

One of the biggest drivers of consumer, business and government behaviour is the current state of the interest rates. Ontario's Ministry of Finance regularly sets tax interest rates for any overdue payments or refunds. This project focuses on predicting quarterly direction of the underpayment interest rate, the overpayment interest rate will be used as an explanatory variable. Analyzing interest rates dynamics will open up new information on inflationary pressure, short-term strategies for economic outcomes and stakeholder insights. Interest rates influence are indirectly represented as signals of economic conditions. They provide additional information to observe fiscal decision patterns. This creates a scenario where interest rate changes can be modeled as a classification problem. Historical analysis will cover quarterly from 2000Q1 to 2026Q1.

Target definition:

r_t = Underpayment Interest Rate in quarter t

\Delta r_t = r_t - r_{t-1}

The classification label is defined as:

- Increase if \Delta r_t > 0
- Decrease if \Delta r_t < 0
- Stable if \Delta r_t = 0

Unit of analysis: One quarter.

Currently, there is no clear quantitative predictive model for interest rates that is able to predict increase, decrease or stable interest rates from current changes and, also, historical data as well (Salem, A. A. M., & Albourawi, A. J. A., 2024).

The absence of a structured and backed up by data classification model is the core issue that the project is focused on. Correct predictions can drastically improve government budgeting and debt management, businesses can rely on the forecast for major changes and improvements, and stakeholders may understand the reasons behind current and future policy changes.

Classification is the most popular analytical theme and the project would be able to classify the direction of policy movement. Using classification and predictive analytics, forecasts can be used to accurately anticipate the direction of interest rates, minimizing the gap in predictive policy analytics.

Research Questions and Justification

First research question: “What is the out-of-sample Macro-F1 score for predicting the quarterly direction of the underpayment interest rate over the period of 2000Q1-2026Q1, using historical rate and macroeconomic indicators?” This topic is evaluating if the problem is predictable using historical and economic data. Out-of-sample performance will reflect real-world predicting rather than in-sample overfitting. Three classes of increase/decrease/stable might not be evenly distributed. This question is directed at the feasibility of whether policymakers can rely on a data-driven model for quarterly changes.

Second research topic: “Do lagged macroeconomic indicators improve Macro-F1 relative to a historical rate model?” Two scenarios can be compared with each other in this equation: historical interest rate features only and historical features with lagged macroeconomic indicators. This topic is looking into the influence of internal dynamics and external economical influences. The main objective is to determine if economic indicators should be added to the forecasting framework.

Third research topic: “How does predictive performance vary over different economic regimes during every eight year period of time, from 2000Q1 to 2026Q1?” This question can focus on crisis and recovery periods of time. If performance stays the same, the structure is predictable and reliable across different economic conditions.

Fourth research topic: “Which predictions best influence classification outcomes as estimated by logistic regression coefficients and SHAP values? What economic signals influence quarterly interest rate changes?” The focus here is in identification of the variables, such as lagged rate, inflation, GDP growth, unemployment, and others, that are strongly correlated with directional changes.

Research questions are directed towards following the project into identifying key predictors of change, evaluation of the feasibility, and further improvement of the model.

Dataset Selection and Rationale

The project uses the dataset of Tax Interest Rates from the Government of Ontario Open Data Portal. The given dataset provides quarterly tax interest rates that are applicable to owed taxes, called underpayments and overpaid taxes, called overpayments. Provincial interest rates are reset quarterly, four times a year (Ontario Ministry of Finance, 2026).

In order to choose the current dataset, it is necessary to notice that it is provided by the official provincial government data website with clear structure and formatting, perfectly fitting to complete the goal of the project (Alice J. Liu and Arpita Mukherjee and Linwei Hu and Jie Chen and Vijayan N. Nair, 2022).

The dataset is directly correlated to the current economic policy developments since the dataset is focused on the first quarter of 2026, from January to March. (Ontario Ministry of Finance, 2026). Potential limitations of the dataset are obviously the temporal scope of the information. This increases the risk of model overfitting and it will require careful validation techniques. Also, factors of economics like inflation and GDP are not included. Moreover, more historical data can be fetched from the official website to answer the last research question and adopt the classification model (*Ontario Economic Accounts*, 2026).

Core Variables: year, quarter, underpayment_rate, overpayment_rate.

Derived Variables: delta_rate, direction, lag1_rate, lag2_rate, rolling_mean_4q, rolling_std_4q.

External indicators: CPI, GDP growth, unemployment rate, Bank of Canada policy rate.

Unit of Analysis: one row represents one calendar quarter.

No existing or future values will be used to prevent data leakage.

Macroeconomic factors will be converted to quarterly frequency, lagged by one quarter.

Proposed Methodology and Tools

Data preprocessing is the first step, when the data is imported, missing values are handled and format is standardized. Then, encoding any categorical features. Feature engineering includes creating prior quarter rates, also known as temporal lags, deriving the variables of change: increase, decrease or unchanged. Application of any external economic factors is completed afterwards. Now, model training can begin.

The skeleton of the model is logistic regression for interpretability. Performance optimization is executed through tree-based models with assessment of stability via cross-validation. The model subsequently can be evaluated for accuracy, precision, recall, F1 score, performance indicator and ROC-AUC, model discrimination ability. Lastly, assess any classification errors and feature effects.

Preprocessing and feature engineering identify core predictors and constructing temporal variables. Model training and comparison evaluate predictive feasibility and the performance of the algorithm. Logistic regression is a baseline of the model that is also needed for interpretability of predictors of the results. Evaluation will use walk-forward (rolling-origin) validation. Training will begin with 2000Q1-2015Q4 and test on 2026Q1. Then, expand forward one quarter at a time.

Baseline models:

1. Always predict “Stable”.
2. Predict the same direction as the previous quarter.

All models must be outperforming the baselines.

All lagged variables and rolling statistics will be determined using past observations. No random shuffling will be used.

Primary Evaluation metric: Macro-F1 score.

Secondary metric: Accuracy, Balanced Accuracy, Confusion Matrix, ROC-AUC.

Python can be mostly operated for the projects, alongside Pandas for the data handling and Numpy for numerical operations. Classification and processing can be done with Scikit-learn, LightGBM, XGBoost. (Chen & Guestrin, 2016). Visualization can be achieved through Matplotlib or Seaborn. Also, Jupyter Notebook could be useful for documentation. All tools are open-source so the project is feasible on standard computing resources (Pensa, R.G., Crombach, A., Peignier, S. et al., 2025).

References

Alice J. Liu and Arpita Mukherjee and Linwei Hu and Jie Chen and Vijayan N. Nair. (2022). *Performance and Interpretability Comparisons of Supervised Machine Learning Algorithms: An Empirical Study*. Cornell University. https://arxiv.org/abs/2204.12868

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. arXiv. https://arxiv.org/abs/1603.02754

Ontario Economic Accounts. (2026). Ontario Data Catalogue. https://data.ontario.ca/dataset/ontario-economic-accounts

Ontario Ministry of Finance. (2026). *Tax interest rates*. Ontario Data Catalogue. https://data.ontario.ca/dataset/tax-interest-rates

Pensa, R.G., Crombach, A., Peignier, S. et al. (2025). *Explaining Random Forest and XGBoost with Shallow Decision Trees by Co-clustering Feature Importance. Mach Learn 114, 287*. Springer Nature Link. https://doi.org/10.1007/s10994-025-06932-9

Salem, A. A. M., & Albourawi, A. J. A. (2024). *Predictive Models for Interest Rate Forecasting Using Machine Learning: A Comparative Analysis and Practical Application.* Brilliance: Research of Artificial Intelligence, 4(2), 764–770. https://doi.org/10.47709/brilliance.v4i2.4983