🔹 Building a Predictive Model for Mortgage Default Risk 📊🚀

💡 **Problem Statement:**With rising interest rates and economic uncertainty, lenders face challenges in predicting mortgage defaults. To help identify and mitigate this risk, I developed a predictive classification model, focusing on macro-economic indicators such as GDP, Mortgage Rates, and Debt Service Ratios to understand how these factors influence delinquency risks.

🔬 What I Did:

✅ Data Collection & Scraping:  
I needed macro-economic data ranging from 2005 to 2024, which posed a significant challenge since many public data sources provided limited time periods (e.g., only 2012 onward). Here’s what I used to collect and prepare the data:

## 1️Sources and Methods:

* 📄 Canadian Economic Indicators (GDP, Unemployment Rate) from StatCan and FRED databases.
* 🏘️ Mortgage Delinquency Rate data using FRED datasets (2005–2024).
* 🌐 Automated data scraping using BeautifulSoup and Requests for regional indicators.
* 📊 Extracted PDF and CSV data manually in cases where automated scrapers failed.

**⚠️ Challenges Faced:**During model building, I encountered overfitting, where the model achieved a 100% accuracy, meaning it was memorizing training data instead of generalizing.

🚩 Here’s what went wrong:

* The model relied too heavily on features with direct correlations (data leakage).
* SMOTE oversampling caused duplicate patterns in synthetic data.
* Strong tree-based models (RandomForest & XGBoost) perfectly fit the training data.

🔍 **How I Overcame Overfitting:**  
1️Stratified Data Splitting: Ensured balanced data distribution in both train and test sets.  
2️Feature Selection: Removed less impactful features (e.g., redundant financial indicators).  
3️Model Simplification: Reduced model complexity using:

* Lower tree depth (max\_depth=4).
* Increased minimum sample requirements (min\_samples\_split=20, min\_samples\_leaf=7).

4Regularization: Applied hyperparameter tuning on RandomForest and XGBoost.  
5️Ensemble Learning: Combined RandomForest, SVM, and Logistic Regression with a Voting Classifier to balance accuracy and generalization.

📊 Key Results and Insights:  
🔵 Final Model Comparison:

* RandomForest: Accuracy = 97.7%
* Logistic Regression: Accuracy = 81.5%
* SVM: Accuracy = 81.1%
* Voting Classifier (Ensemble): Accuracy = 83.7% (balanced generalization)

📌 Feature Importance Highlights:

* 🔴 Unemployment Rate (35%) → Strongest driver of delinquency risk.
* 🟠 Mortgage Rates (30%) → Rising interest rates increase payment pressure.
* 🟡 Debt Service Ratio (17%) → Indicates financial stress.
* 🟢 Dataflow (18%) → Reflects household financial conditions.

**🚀 Next Steps**:  
💡 Deploy as an API or Web App (Flask or Streamlit).  
💡 Incorporate Borrower-Level Data: Loan-to-Value (LTV), Debt-to-Income (DTI), Credit Scores.  
💡 Test and Validate: Apply model to real-world Canadian mortgage data and monitor its performance.

📢 Lessons Learned:  
This project taught me the importance of data sourcing, resampling challenges, and model evaluation techniques. Managing data leakage and feature redundancy was critical in achieving better generalization and avoiding overfitting.

💼 Looking Ahead:  
Would love to hear feedback from others working in AI for Finance or those tackling macroeconomic risk predictions! If you have suggestions or insights into borrower-level data sources, feel free to share. Let’s collaborate and learn together! 🚀

🔗 Key Tags:  
#AI #MachineLearning #DataScience #PredictiveModeling #MortgageRisk #Macroeconomics #FinTech #DataEngineering #RiskManagement #AIforFinance