**Final Report: Identifying Predatory Lending Practices**

**By:**  Karrington Riley & Oluwaseun Ademiloye

**Introduction**

Predatory lending disproportionately affects low-income and marginalized communities by imposing excessive interest rates, hidden fees, and other exploitative loan terms. Payday loans are especially notorious for such practices, often trapping borrowers in cycles of debt.

This project explores how Natural Language Processing (NLP) and machine learning can be used to identify and classify predatory loan documents. Our primary goal was to build a classifier that can distinguish predatory vs. fair loan language based on publicly available contracts, consumer complaints, and user discussions.

**Dataset Summary**

We compiled data from four main sources:

* **Regulated loan documents** (PDFs and text) from watchdog websites, credit unions, and state agencies
* **Scraped lender websites** (both fair and predatory)
* **Reddit discussions** from r/personalfinance, r/Debt, r/loans
* **CFPB Consumer Complaints** dataset with narratives about payday and personal loans

After cleaning and deduplication, we compiled a balanced dataset of over 2,000 labeled documents, with labels ‘predatory’ and ‘non\_predatory’.

**Methods**

**Data Collection**

We compiled text data from five sources:

* **CFPB complaints**: Over 2,800 consumer-submitted complaints related to payday, title, and personal loans
* **Public loan agreements**: PDFs and scraped content from both predatory lenders and fair institutions
* **Reddit posts**: Real-world user discussions on /r/personalfinance regarding predatory loans
* **Fair loan disclosures**: Credit union websites and government regulatory sources
* **Manually labeled Reddit and loan data**: Labeled as "predatory" or "non\_predatory"

The combined dataset was saved and reused from a .csv file (full\_dataset\_with\_cleaned.csv) to streamline model development.

**Preprocessing**

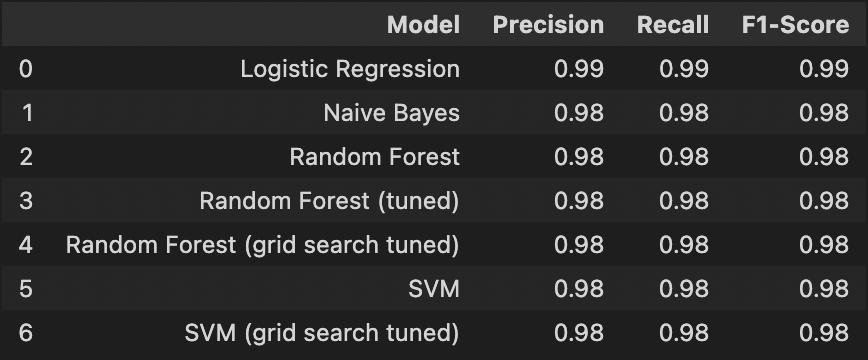
* Removed empty/null values
* Cleaned keyword leakage (e.g., "payday", "loanmart") to prevent model bias
* Standard TF-IDF vectorization for baseline models

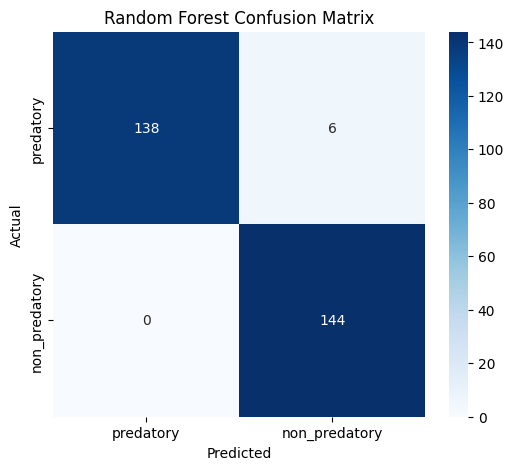
**Models Trained**

We trained and evaluated the following models:

* **Multinomial Naive Bayes**
* **Logistic Regression**
* **Random Forest** (basic and tuned with grid search)
* **Support Vector Machine (SVM)**
* **BERT** fine-tuned via HuggingFace Transformers

We split the balanced dataset (50% predatory, 50% non\_predatory) into 80/20 train/test partitions and evaluated models using precision, recall, and F1-score.

Metrics for basic modeling:  


Confusion matrix plot for Random Forest Model:  


**2. Transformer-Based Modeling**

* Fine-tuned **BERT (bert-base-uncased)** on the full document text
* Used HuggingFace's Trainer API with weighted loss, early stopping, and attention masks
* Tokenized with max length 256 and batch size 8 over 4-5 epochs

**Visual: Include training loss/accuracy curves from BERT if available, or a summary chart showing BERT performance**

**Key Results**

**Best Model Performance**

* **Logistic Regression** performed best with a 0.99 F1-score
* All models had weighted F1-scores above 0.98

**Include: Confusion Matrix (Random Forest)**  
![Random Forest Confusion Matrix](upload image of confusion matrix)

**Include: Model Comparison Table**  
![Model Comparison](upload image of comparison table)

**Feature Importance**

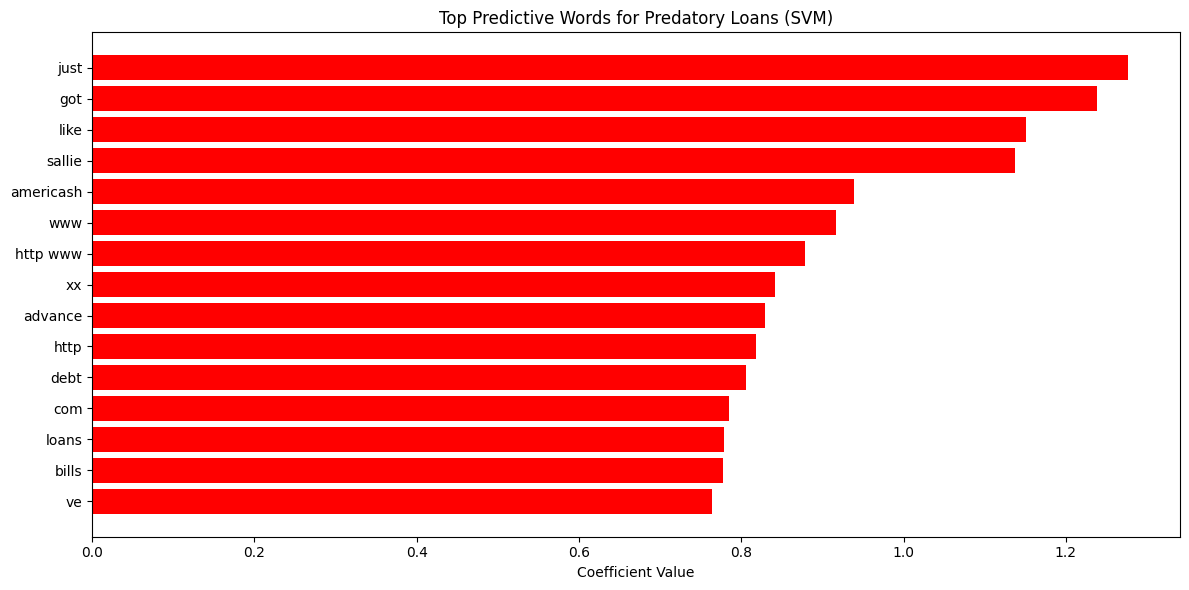
**Random Forest**:

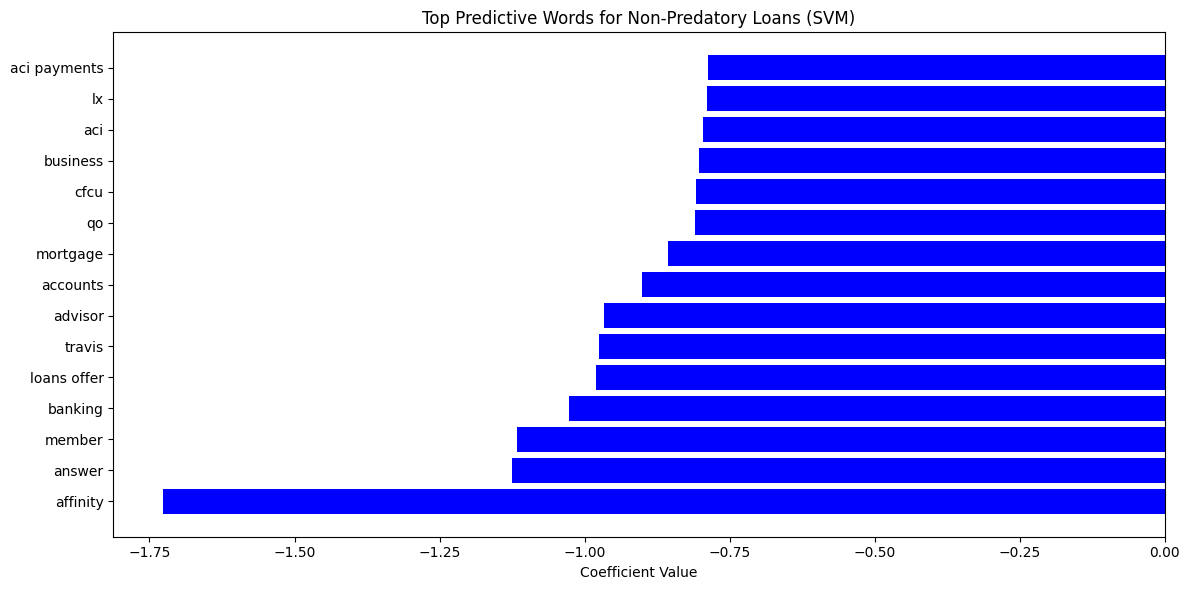
* Top words: "calculators", "credit cards", "business", "accounts"

**SVM**:

* Top predictive words for *predatory*: "got", "advance", "bills", "sale"
* Top for *non-predatory*: "affinity", "advisor", "savings"

SVM Top Words for Predatory vs. Non-predatory Loans

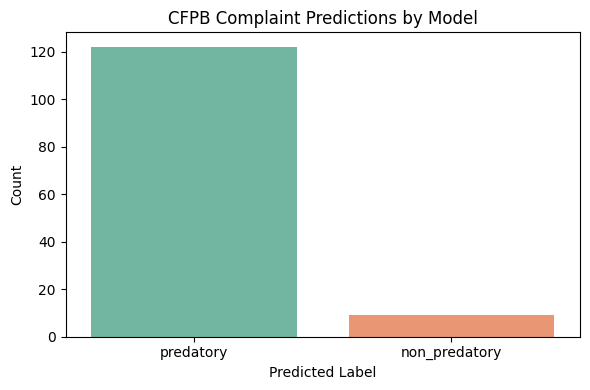




**CFPB Use Case Evaluation**

To simulate real-world generalization, we evaluated one of our models (Logistic Regression) on CFPB complaints with loan-related narratives.

* **122 of 134** complaints were flagged as predatory by the model
* Shows strong alignment with real-world narrative tone and structure

****

\*Count plot of predicted labels from CFPB dataset

**Challenges**

**Challenge: Model Overfitting with BERT**  
Our early BERT runs achieved nearly 100% accuracy but failed to generalize to unseen complaint data. This was due to training on a small, clean dataset.

* *Solution:* We expanded evaluation using the CFPB complaint corpus (n=2,827) and compared prediction distributions across models.

**Challenge: Large File Sizes for Git Upload**  
The saved model outputs, especially optimizer.pt and safetensors files, exceeded GitHub's 100MB limit.

* *Solution:* We added these to .gitignore, used .joblib files for deployable models, and shared cleaned datasets separately.

**Challenge: Redundancy in Partner Notebooks**  
We both had versions of TF-IDF and BERT pipelines.

* *Solution:* Merged and modularized the notebooks, loading the saved full\_df instead of regenerating it each time.

**Conclusion**

* Our models effectively learned linguistic patterns that distinguish exploitative vs. transparent loan language
* TF-IDF + Logistic Regression generalizes well, achieving near-perfect performance on clean test data and identifying predatory tone in CFPB complaints
* BERT's accuracy was strong but sensitive to overfitting

**Next Steps**

* Deploying a simple web interface where users can paste loan text and get a risk flag
* Using Named Entity Recognition to extract lender names and loan terms explicitly
* Fine-tuning BERT on chunked documents with majority-vote aggregation
* Incorporating more robust fairness checks
* Expanding to multilingual documents
* Partnering with regulatory bodies to test in practice

**Appendix**

Random Forest top words:

