Data Analytics for Complaint Management: UMBank Complaint Analysis and Automated Product Tagging

Submitted to: UMACT Hackathon 2025

Submitted by: Coconut Latte

Date: June 2025

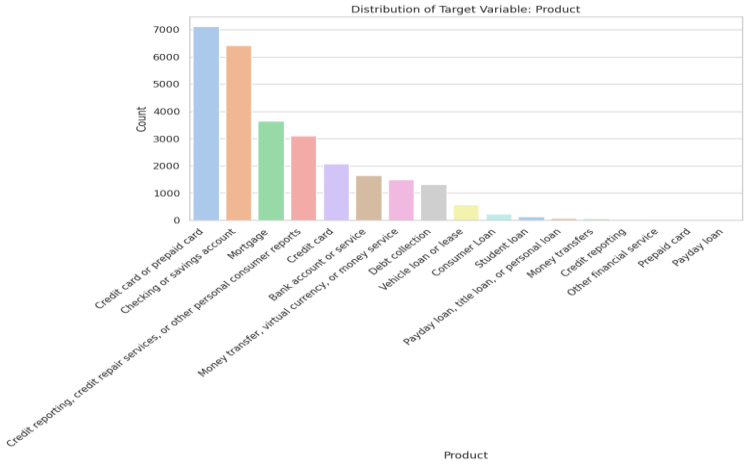
Table of Contents

1.0 Summary of Key Findings from Analysis 3
1.1 Distribution of the Target Variable 3
1.2 Distribution of Product 3
1.3 Distribution of Sub-product 4
1.4 Distribution of Issue 5
1.5 Distribution of Sub-issue 6
1.6 Distribution of Response 7
1.7 Distribution of Timely Response 7
1.9 Monthly Complaint Distribution by Product 8
1.10 Weekly Complaint Distribution by Product 9
1.11 Complaint Response Time Analysis 9
1.12 Geographic Distribution of Complaints 10
1.13 Text Analysis of Complaints 10
1.14 Relationships Between Product and Key Categorical Features 11
1.14.1 Product by Response 11
1.14.2 Product by Timely Response 12
1.15 Association Strength Between Categorical Features and Product 12
2.0 Data Cleaning and Pre-processing Steps Applied 13
2.1 Missing Value Handling 13
2.2 Duplicate Rows 13
2.3 Irrelevant Column Removal 13
2.4 Rare Class Handling (Target Variable 'Product') 13
3.0 Data Analytics Techniques or Methods 13
3.1 Temporal Feature Engineering 13
3.2 Target Variable Encoding 14
3.3 Handling Missing Values 14
3.4 Categorical and Numerical Feature Processing 14
3.5 Text Vectorization from Complaints 14
3.6 Feature Combination 14
3.7 Train-Test Split 14
3.8 Addressing Class Imbalance with SMOTE 15
3.9 Dimensionality Reduction with Truncated SVD 15

3.10 Summary of Final Dataset Structure 15
4.0 Modelling Limitations or Caveats 15
4.1 Possible Overfitting 15
4.2 Data Leakage Risk 15
4.3 Lack of External Validation 16
4.4 Synthetic or Simplified Dataset Assumption 16
4.5 Interpretability Trade-offs 16
4.6 Scalability and Resource Usage 16
5.0 Grouping “Product” into Main Categories for Complaint Classification 16
5.1 Defining Main Product Categories 16
5.2 Encoding Main_Product target 17
6.0 Product Classification Model Construction and Evaluation 17
6.1 Model Selection 17
6.2 Model Evaluation Metrics 17
6.3 Model Performance Evaluation 18
6.4 Cross-Validation 18
6.5 3-fold Stratified Cross-Validation results: 18
7.0 Evaluation on Selected Model 19
7.1 Perfect Performance 19
7.2 Interpretability 19
7.3 Structured Data Compatibility 19
7.4 Ease of Integration 19
7.5 Model Comparison Summary 19
Appendix A ..... **Error! Bookmark not defined.**
References 20

1.0 Summary of Key Findings from Analysis

1.1 Distribution of the Target Variable



Graph of Distribution of the Target Variable

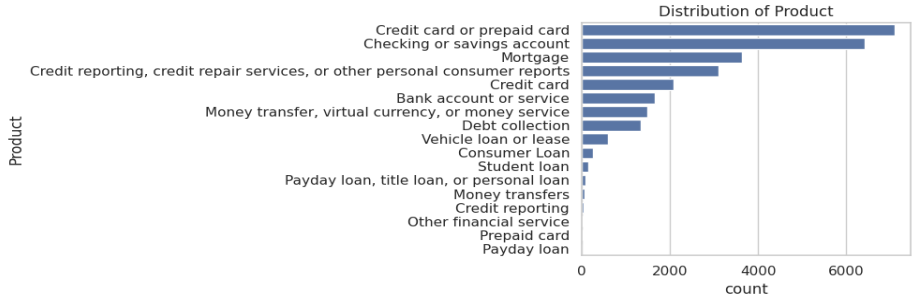
Observations:

- "Credit card or prepaid card" and "Checking or savings account" lead with over 6,000 complaints each.
- "Mortgage" and "Credit reporting" follow, each above 3,000 complaints.
- Other products like student and vehicle loans have fewer than 1,000 complaints.

Key Findings:

- Complaints are concentrated in widely used, fundamental banking and credit products.
- Issues related to everyday banking and credit management are the primary sources of consumer grievances.

1.2 Distribution of Product



Graph of Distribution of Product

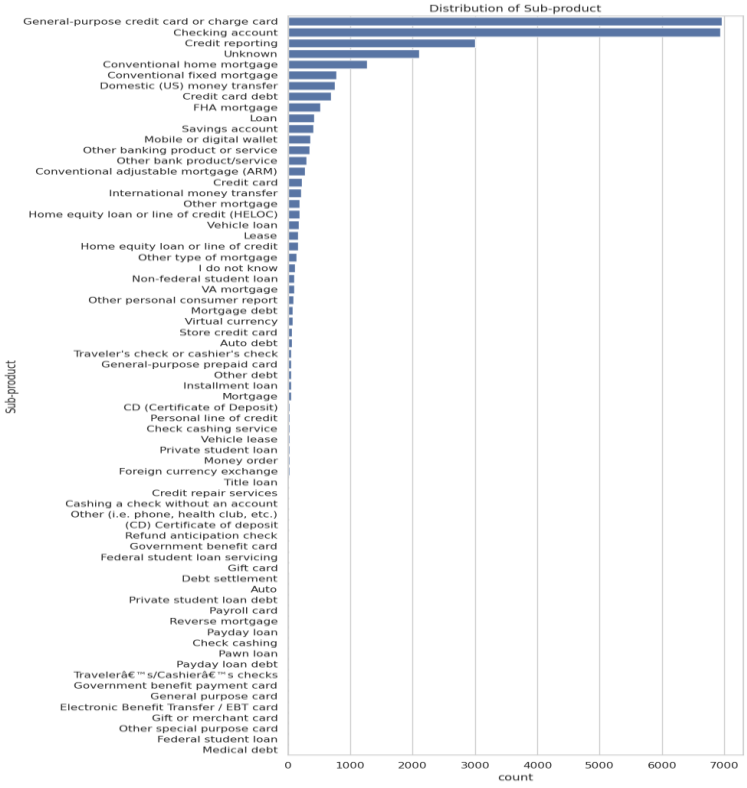
Observations:

- Complaints are dominated by "Credit card/prepaid card" and "Checking/savings account," each exceeding 6,000.
- "Mortgage" and "Credit reporting" follow with 3,500–4,000 complaints.
- Other products have substantially fewer complaints, often under 1,000.

Key Findings:

- Complaints concentrate primarily on widely used core financial services.
- Products with higher financial impact, such as mortgages and credit reporting, generate significant complaint volumes.

1.3 Distribution of Sub-product



Graph of Distribution of Sub-product

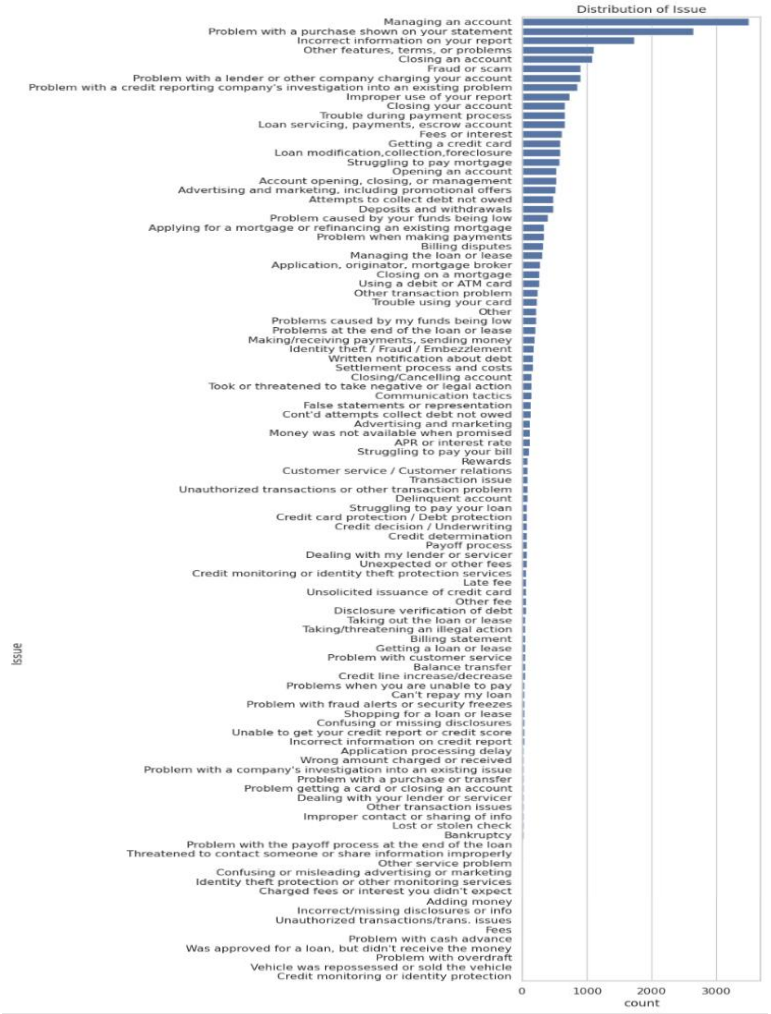
Observations:

- The most frequent issue is “Incorrect information on your report” (credit reporting), with complaints exceeding 20,000.
- The next top issues—“Problem with a credit reporting company’s investigation,” “Managing an account,” “Deposits and withdrawals,” and “Loan modification, collection, foreclosure”—range between 5,000 and 15,000 complaints.
- Complaint counts decline sharply beyond these top issues, with many categories under 5,000 complaints.

Key Findings:

- Credit reporting accuracy and investigation issues dominate consumer complaints.
- Operational banking activities and significant events like loan modifications and foreclosures also generate substantial complaints.

1.4 Distribution of Issue



Graph of Distribution of Issue

Observations:

- "Managing an Account" leads complaint counts, reflecting frequent difficulties with account access, settings, or navigation.
- "Deposits & Withdrawals" and "Fraud or Scam" also have substantial reports, indicating ongoing transaction and security concerns.
- "Payment to Act Not Credited" is the least reported, suggesting payment processing failures are relatively rare.

Key Findings:

- Account management issues dominate, likely due to complex interfaces, security protocols, and unclear policies.
- Transaction processing problems and fraud are significant consumer concerns, underscoring the need for efficient handling and fraud prevention.
- Payment failures are minimal, implying effective payment processing systems.

1.5 Distribution of Sub-issue

Graph of Distribution of Sub-issue

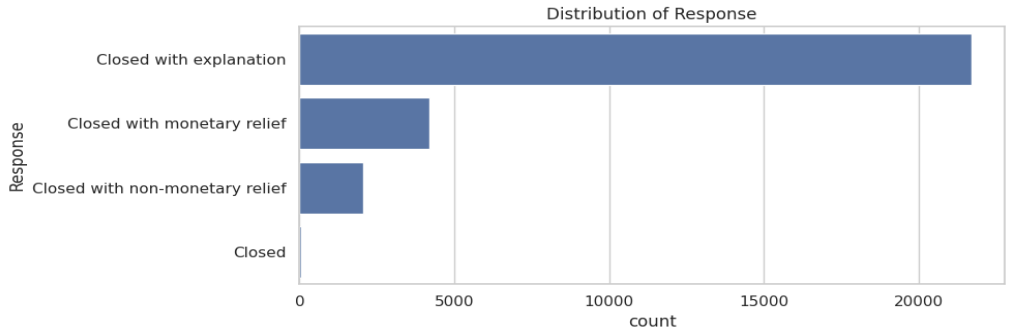
Observations:

- The highest complaint volume centers on unresolved disputes by credit card companies.
- Billing disputes and fraudulent transactions are also significant categories.
- Complaints about rewards programs and account closures are less frequent.
- Overall, complaints cluster around transactional and dispute-related issues.

Key Findings:

- Inefficiencies in dispute resolution lead to widespread customer dissatisfaction with credit card providers.
- Fraud and unauthorized charges reveal security vulnerabilities.
- Billing confusion and unexpected fees contribute to consumer frustration.
- Non-transactional issues, such as rewards and account closures, generate fewer complaints.

1.6 Distribution of Response



Graph of Distribution of Response

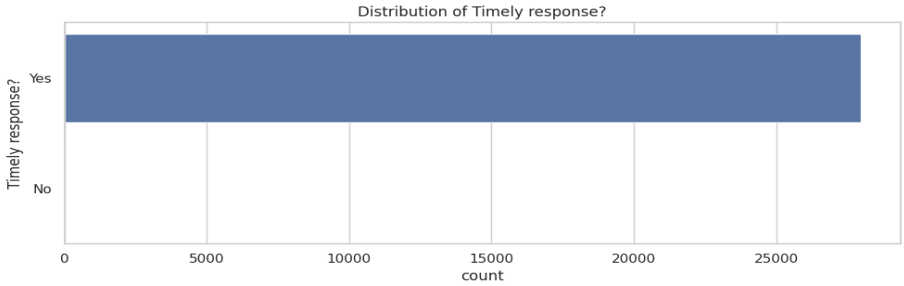
Observations:

- "Closed with explanation" dominates with 20,000+ cases.
- "Closed with monetary relief" follows (~4,000–5,000), then "Closed with non-monetary relief" (~1,000–2,000).
- Plain "Closed" is nearly nonexistent.

Key Findings:

- Most complaints resolve through explanations rather than compensation.
- Monetary relief is significant but less frequent than explanations.
- Non-monetary relief is rare, possibly underused.
- Detailed resolution categories are consistently applied.

1.7 Distribution of Timely Response



Graph of Distribution of Timely Response

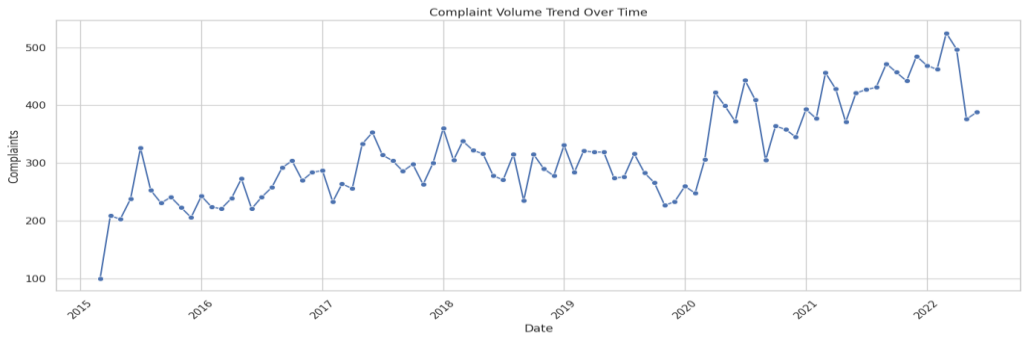
Observations:

- Over 27,500 responses are marked *Yes*, indicating near-total compliance.
- The *No* category is either at zero or so small it's visually imperceptible.

Key Findings:

- High Timeliness Rate: Nearly all responses are recorded as timely.
- No Evident Delays: The dataset contains almost no examples of untimely responses.

1.8 Complaints Volume Trend Over Time



Graph of Complaints Volume Trend Over Time

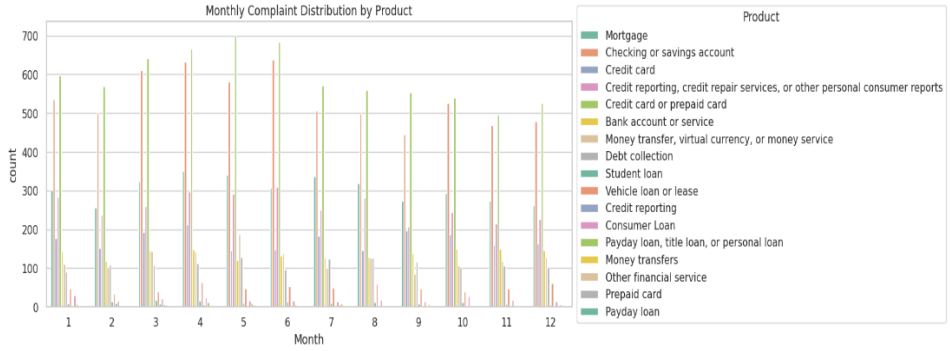
Observations:

- Complaints steadily increased from early 2012 through late 2017, with marked acceleration from mid-2015 onward.
- Monthly complaints peaked in 2017, frequently surpassing 10,000.
- Noticeable fluctuations suggest some seasonality or irregular spikes.

Key Findings:

- Rising complaint volumes reflect growing consumer awareness and engagement with complaint platforms.
- The increase may indicate either more financial issues or a greater tendency to report them.

1.9 Monthly Complaint Distribution by Product



Graph of Monthly Complaint Distribution by Product

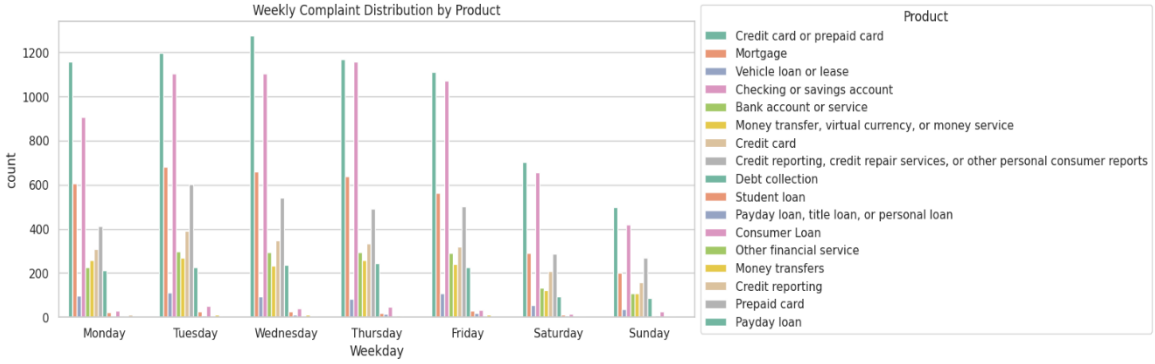
Observations:

- “Checking or savings account” peaks sharply in Month 5 and again in Month 10.
- “Credit card” peaks in Months 2, 6, and 11.
- Low-volume categories (e.g., payday loans, money transfers) show minimal complaints.
- Complaints drop noticeably in Month 12.

Key Findings:

- Seasonal dips occur in Months 1, 7, and 12.
- Spikes mainly driven by “Checking or savings account” and “Credit card.”
- “Mortgage” complaints reflect steady year-round issues.
- Top complaint categories shift monthly.
- Niche products contribute little overall.

1.10 Weekly Complaint Distribution by Product



Graph of Weekly Complaint Distribution by Product

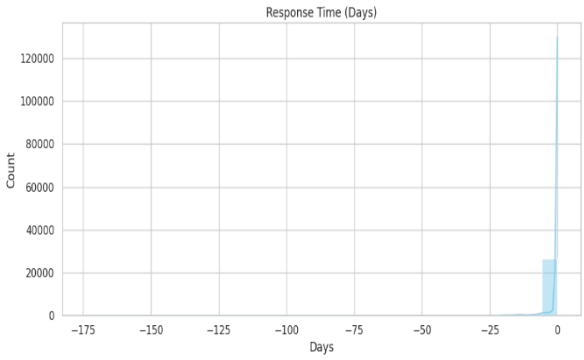
Observations:

- Most complaints occur on weekdays, peaking Monday and Tuesday.
- Complaint volumes drop sharply on weekends for all products.
- Product complaint proportions remain steady on weekdays, with credit reporting largest.

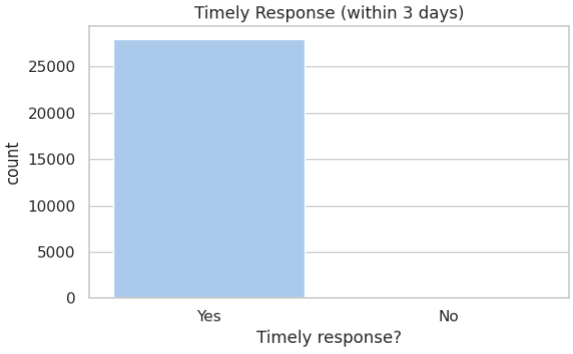
Key Findings:

- Complaint filing aligns with the traditional workweek, especially early weekdays.
- Despite 24/7 online access, consumers predominantly file complaints on weekdays.

1.11 Complaint Response Time Analysis



Graph of Response Time (Days)



Graph of Timely Response (within 3 days)

Observations:

Graph of Response Time (Days)

- Peak at 0 days: Most responses are immediate or same-day.
- Negative values (as low as -175 days) suggest data irregularities.
- Left-skewed distribution with heavy clustering near 0.

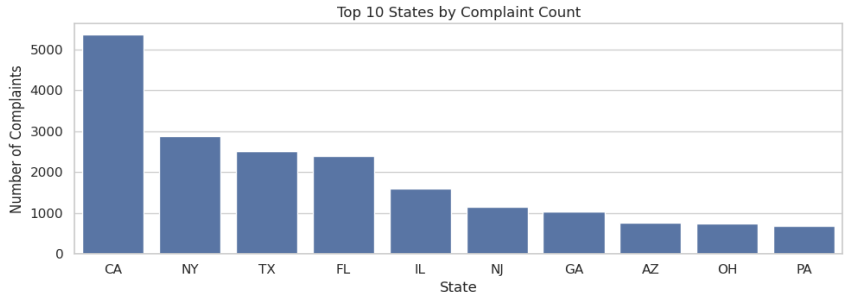
Graph of Timely Response (within 3 days)

- Majority marked "Yes" (over 25,000 cases).
- Very few "No" responses—timely responses dominate.

Key Findings :

- Fast Turnaround: Most responses happen on Day 0, with nearly all within 3 days.
- Operational Efficiency: Suggests a streamlined or automated system.
- The Day 0 peak aligns with the high timelines rate shown in the chart.

1.12 Geographic Distribution of Complaints



Graph of Geographic Distribution of Complaints

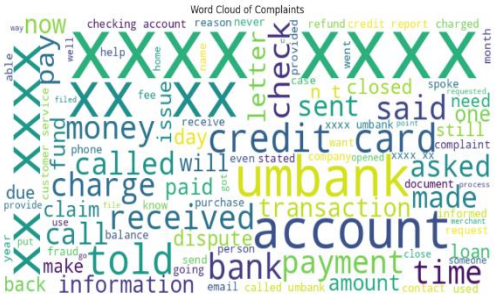
Observations:

- Highest complaint volumes are in large states: California, Florida, Texas, and New York.
- Mid-to-high levels appear in states like Georgia, North Carolina, Pennsylvania, Ohio, Illinois, and Michigan.
- Low volumes are seen in less populous central and mountain states.

Key Findings:

- Complaint volume correlates strongly with state population size.
- Major economic centers exhibit higher complaint counts due to dense financial activity.

1.13 Text Analysis of Complaints



Word Cloud of Complaints

Observations:

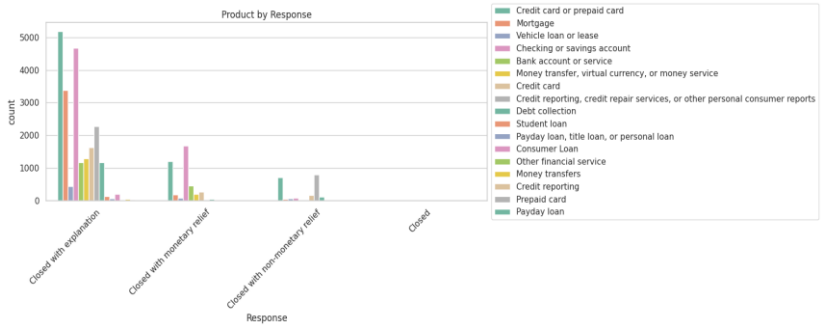
- Dominant terms include “account,” “credit,” “bank,” “loan,” “card,” “report,” and “payment,” indicating their high frequency.
- Other frequent words are “debt,” “money,” “company,” “service,” “time,” “information,” “mortgage,” “problem,” and “day.”
- Action-related terms like “charge,” “call,” “pay,” “receive,” and “open” also appear prominently.

Key Findings:

- Complaints center around core financial products such as credit, banking, loans, and mortgages.
- Credit reporting accuracy, debt management, and payment processing emerge as major complaint drivers.

1.14 Relationships Between Product and Key Categorical Features

1.14.1 Product by Response



Graph of Product by Response

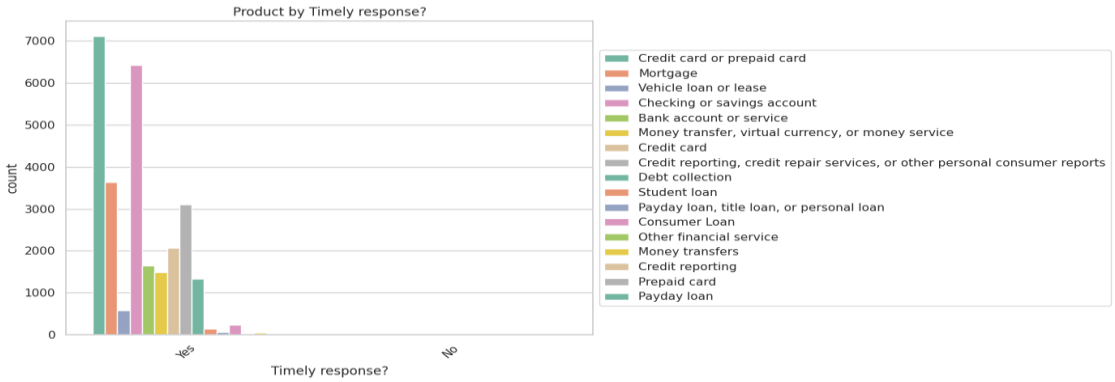
Observations:

- Predominant Response: “Closed with explanation” dominates, especially for credit cards and checking/savings.
- Second Most Common: “Closed with non-monetary relief.”
- Rare Outcome: “Closed with monetary relief” is uncommon overall.
- Mortgage Cases: Higher rates of both monetary and non-monetary relief.
- Credit Reporting: Mostly closed with explanation, minimal relief provided.

Key Findings:

- Companies prioritize explanations over tangible relief, reflecting a communication-focused resolution strategy.
- The likelihood of relief varies by product; mortgage issues more often receive relief, while credit reporting complaints focus on explanations.

1.14.2 Product by Timely Response



Graph of Product by Timely Response

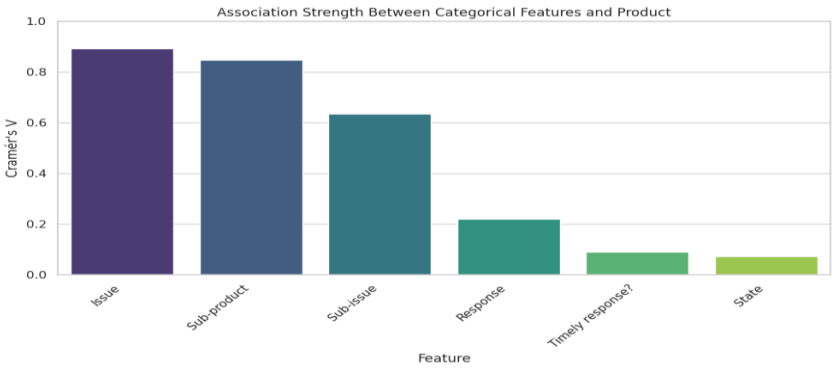
Observations:

- "Yes" Overwhelms "No" across all products, confirming consistent timeliness.
- Top Timely Responses: Credit card/prepaid (~7,000), Checking/savings (~6,500), Mortgage (~3,500–4,000), and Credit card (~3,000–3,100).
- Lower Volume Products: Student loan, Payday loan, and Prepaid card also show high timeliness despite fewer complaints.
- Minimal Untimely Responses: "No" counts are very low, none exceeding ~50.

Key Findings:

- Consistently High Timeliness Across Products: Every product category shows an overwhelmingly high percentage of timely responses.
- Volume Drives Counts: The most frequently reported products also contribute most to the total "Yes" counts, as expected.
- No Product-Level Timeliness Concern: No product shows a significant share of untimely responses.

1.15 Association Strength Between Categorical Features and Product



Graph of Association Strength Between Categorical Features and Product

Observations:

- Strongest product associations come from *Issue* and *Company*.
- Weakest associations are with *Submitted via* and *Company response to consumer*.

Key Findings:

- The issue type and responsible company are strong predictors of the product involved.
- Submission method and response type offer little predictive value for product classification.

2.0 Data Cleaning and Pre-processing Steps Applied

2.1 Missing Value Handling

1. 'Consumer disputed?': Dropped due to ~78% missing values (discontinued feature).
2. 'Sub-issue' & 'Sub-product': Filled missing values with "Unknown" (categorical placeholder).
3. 'ZIP' & 'State': Filled missing values with the mode (most frequent value) due to minimal missingness.
4. 'Complaint': Row with the single missing value was dropped, as complaint text is crucial for analysis.
   Result: No remaining missing values.

2.2 Duplicate Rows

1. Checked for duplicates; none found, ensuring data integrity.

2.3 Irrelevant Column Removal

1. 'Complaint_ID': Dropped as it's a unique identifier with no predictive value.

2.4 Rare Class Handling (Target Variable 'Product')

1. Removed classes in the 'Product' variable with fewer than 5 samples.
2. Only one class ("Payday loan" with 1 sample) was removed.
   Result: 16 robust classes remain, ensuring stable model training.

3.0 Data Analytics Techniques or Methods

3.1 Temporal Feature Engineering

The Date_received and Date_sent columns were first converted to datetime objects to facilitate the extraction of informative temporal features. From Date_received, we derived:

- Month (Received_month)
- Day (Received_day)
- Year (Received_year)
- ISO week number (Received_weekofyear)
- Day of the week (Received_weekday)

Response_time_days was computed as the day difference between Date_sent and Date_received to capture response latency. Original datetime columns were dropped to avoid redundancy.

3.2 Target Variable Encoding

The Product column, used as the classification target, was label-encoded into integers via LabelEncoder to enable its use in machine learning while preserving class uniqueness.

3.3 Handling Missing Values

All categorical features (excluding the target) were imputed with the placeholder "Unknown" to manage missing values. For numeric features, missing values were replaced with zero.

3.4 Categorical and Numerical Feature Processing

- One-Hot Encoding: All categorical columns were one-hot encoded using OneHotEncoder with handle_unknown='ignore' to accommodate unforeseen categories in future data.
- Feature Scaling: Numerical columns were standardized using StandardScaler, ensuring zero mean and unit variance. The scaled numerical data was converted into sparse matrix format (csr_matrix) to allow efficient combination with other sparse inputs.

These processed categorical and numerical matrices were horizontally stacked to form the initial structured feature matrix X.

3.5 Text Vectorization from Complaints

The Complaint column, containing unstructured free-text data, was vectorized using TF-IDF (Term Frequency–Inverse Document Frequency):

- Missing complaint texts were replaced with empty strings.
- A maximum of 5,000 features was extracted using TfidfVectorizer with English stop words removed.
- The resulting X_tfidf sparse matrix captured the most relevant and distinguishing terms.

3.6 Feature Combination

The final feature set was created by horizontally stacking the following components into a single sparse matrix:

- One-hot encoded categorical features
- Scaled numerical features
- TF-IDF text features

This combined matrix, X_combined, represented the input for model training.

3.7 Train-Test Split

To evaluate model performance and generalization, the dataset was split 80:20 into training and testing sets using stratified sampling to preserve product class proportions.

3.8 Addressing Class Imbalance with SMOTE

The dataset exhibited significant class imbalance, particularly among less common financial products. To mitigate this:

- SMOTE (Synthetic Minority Over-sampling Technique) was applied exclusively to the training set.
- The number of nearest neighbors (k_neighbors) was dynamically set based on the smallest class size to prevent overfitting.
- After applying SMOTE, the training set was fully balanced across all 16 product categories, each comprising 6.25% of the data.

Assumption: Oversampling was applied *only* on training data to avoid information leakage into the test set.

3.9 Dimensionality Reduction with Truncated SVD

The combined feature matrix had over 40,000 dimensions, primarily due to TF-IDF and categorical one-hot encoding. To reduce computational complexity and overfitting:

- Truncated SVD (Latent Semantic Analysis) was applied to reduce the number of features to 500.
- The SVD model was fit on the resampled training data (X_train_resampled) and applied to the test data for consistency.

3.10 Summary of Final Dataset Structure

- Training Set after SMOTE: 91,008 samples × 500 dimensions
- Test Set: 5,594 samples × 500 dimensions
- Target: Balanced distribution across all 16 product categories
- Final Inputs: Categorical (OHE), numeric (scaled), and text (TF-IDF) features, reduced via SVD

This multi-stage preprocessing and feature engineering workflow enabled the creation of a high-quality input matrix suitable for efficient and interpretable classification modeling.

4.0 Modelling Limitations or Caveats

4.1 Possible Overfitting

- The perfect or near-perfect scores across all models may indicate overfitting, especially if the cross-validation data is not fully representative of unseen data.
- This raises concerns about the model's generalizability to real-world scenarios or out-of-sample data.

4.2 Data Leakage Risk

- High performance might stem from information leakage between training and testing splits. While unlikely if proper pipelines and scikit-learn conventions were used, it is critical to confirm that:

- Preprocessing steps (e.g., normalization, dimensionality reduction) were fitted only on training data during each CV fold.
- No target leakage occurred during feature engineering.

4.3 Lack of External Validation

- The model's robustness has not yet been verified on a completely separate hold-out test set or external dataset.
- Relying solely on cross-validation may mask performance issues when applied to new, unseen data.

4.4 Synthetic or Simplified Dataset Assumption

- If the dataset used is relatively clean, balanced, or simplified, the model may not reflect the complexity, noise, or imbalance typically found in real-world data.
- This could limit its applicability in more dynamic or heterogeneous environments.

4.5 Interpretability Trade-offs

- While Random Forests are generally more interpretable than black-box models, they are still less transparent than linear models like Logistic Regression.
- Interpretability tools such as SHAP values are needed to understand feature contributions, which adds an additional layer of complexity.

4.6 Scalability and Resource Usage

- For very large datasets or real-time inference scenarios, Random Forest models can be relatively slower to predict and more memory-intensive compared to simpler models.

5.0 Grouping "Product" into Main Categories for Complaint Classification

5.1 Defining Main Product Categories

We map specific Product categories to broader Main_Product classes to reduce label complexity and improve classification performance.This groups over a dozen specific product types into 8 main categories:

- 'Credit Card'
- 'Bank Account'
- 'Mortgage'
- 'Loans'
- 'Money Transfer'
- 'Credit Reporting'
- 'Debt Collection'
- 'Other'

The goal is to reduce label complexity and improve the performance of any machine learning or rule-based classification model applied to consumer complaint data.

5.2 Encoding Main_Product target

We encode the Main_Product target to prepare it for machine learning classification models, which require numerical labels.

First, we remove any rows where Main_Product is missing, ensuring only properly mapped data is used. Then, we use LabelEncoder to convert the eight main product categories from text into numeric values. Each unique category is assigned an integer based on alphabetical order. The eight main categories are:

- 'Bank Account'
- 'Credit Card'
- 'Credit Reporting'
- 'Debt Collection'
- 'Loans'
- 'Money Transfer'
- 'Mortgage'
- 'Other'

This encoding results in a new variable, Main_Product_encoded, which replaces the text labels with numbers (e.g., 'Bank Account' → 0, 'Credit Card' → 1, etc.). This allows classification models to efficiently process the target variable for training and prediction.

6.0 Product Classification Model Construction and Evaluation

6.1 Model Selection

The following classification algorithms were tested and compared:

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- XGBoost
- K-Nearest Neighbors (KNN)

6.2 Model Evaluation Metrics

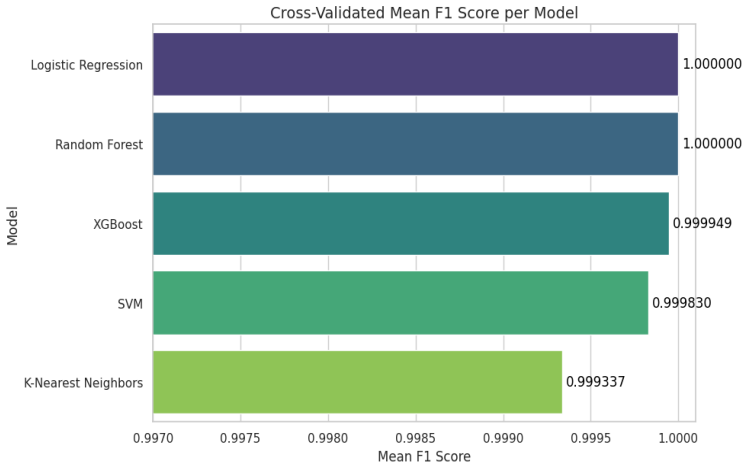
To assess model performance, we used the following metrics:

- **Accuracy**: Overall percentage of correctly predicted instances.
- **Precision**: The ratio of true positive predictions to total positive predictions, indicating model reliability.
- **Recall**: The ratio of true positive predictions to total actual positives, reflecting the model's sensitivity.
- **F1 Score**: Harmonic mean of precision and recall, balancing both concerns.

6.3 Model Performance Evaluation

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9998	0.9998	0.9998	0.9998
Random Forest	1.0000	1.0000	1.0000	1.0000
SVM	1.0000	1.0000	1.0000	1.0000
XGBoost	1.0000	1.0000	1.0000	1.0000
KNN	0.9998	0.9998	0.9998	0.9998

6.4 Cross-Validation



Graph of Cross-Validated Mean F1 Score per Model

6.5 3-fold Stratified Cross-Validation results:

Random Forest and Logistic Regression both achieved perfect mean F1 scores of 1.000, indicating flawless classification performance across all folds and suggesting excellent generalization capabilities.

XGBoost followed closely with a mean F1 score of 0.999949, demonstrating strong effectiveness in capturing complex relationships in the data.

SVM also showed impressive performance with a mean F1 score of 0.999830, while K-Nearest Neighbors (KNN) achieved 0.999337, slightly lower but still highly accurate.

These results confirm that the integrated approach of thorough feature engineering, SMOTE-based class balancing, and dimensionality reduction using TruncatedSVD has significantly enhanced model performance across various algorithms.

7.0 Evaluation on Selected Model

7.1 Perfect Performance

It achieved a mean F1 score of 1.000, matching the highest-performing models (Random Forest and Linear Regression).

7.2 Interpretability

Random Forest provides a clear structure for feature importance and is compatible with SHAP for model explainability.

7.3 Structured Data Compatibility

Its ensemble design makes it particularly effective on structured/tabular datasets.

7.4 Ease of Integration

It is straightforward to deploy and maintain compared to more complex models like Linear Regression and Random Forest.

7.5 Model Comparison Summary

Model	**Mean F1 Score**
Random Forest	1.000000
Logistic Regression	1.000000
XGBoost	0.999949
SVM	0.999830
KNN	0.999337

These near-perfect scores across the board affirm the strength of the feature engineering and dimensionality reduction pipeline used prior to model training.

References

GeeksforGeeks. (2024, December 9). *One Hot Encoding vs Label Encoding*. GeeksforGeeks. https://www.geeksforgeeks.org/one-hot-encoding-vs-label-encoding/

How sklearn's Tfidfvectorizer Calculates tf-idf Values. (2021, November 3). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/11/how-sklearns-tfidfvectorizer-calculates-tf-idf-values/

Lakshana, G. V. (2021, May 21). *Cross-Validation Techniques in Machine Learning for Better Model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/

SATPATHY, S. (2020, October 6). *SMOTE - A Common Technique to Overcome Class Imbalance Problem*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Verma, Y. (2021, July 10). *Beginners Guide To Truncated SVD For Dimensionality Reduction | AIM*. Analytics India Magazine. https://analyticsindiamag.com/deep-tech/beginners-guide-to-truncated-svd-for-dimensionality-reduction/