Medical Claim Denial Prediction Using Machine Learning

Aim:

Medical Billing Analyst at a Revenue Cycle Management (RCM) company, responsible for analyzing claim patterns, identifying denial trends, and improving collections through data-driven insights.

Business Objective:

The goal of this project is to proactively predict whether a medical claim is likely to be denied based on key billing details. By leveraging machine learning techniques, the model uses historical data including CPT codes, insurance company names, physician names, and billed amounts to classify claims as either "Denied" or "Not Denied." This enables teams to intervene before submission, correct potential errors, and improve first-pass resolution rates.

Why It Matters:

Claim denials are a major cause of revenue leakage for healthcare providers. This solution helps reduce manual effort in reviewing and reworking denials by:

- Identifying high-risk claims upfront.
- Providing insights into patterns behind denials.

1)Identify Top Denied CPT Codes

A. Rank CPTs by frequency of denial or non-payment

Rank	CPT Code	Denial Count
1	11721	89
2	29580	88
3	81001	86
4	74177	83
5	90471	83
6	99285	81
7	99232	80
8	20550	78
9	99214	78
10	99203	78

B. Show denial rates per CPT

Rank	CPT Code	Total Claims	Denied Claims	Denial Rate (%)
1	99285	248	92	37.10%
2	11721	258	90	34.88%
3	93000	248	82	33.06%
4	87635	278	91	32.73%
5	20610	250	81	32.40%
6	29580	263	84	31.94%
7	99214	248	77	31.05%
8	20550	249	77	30.92%
9	99291	270	83	30.74%
10	74177	238	73	30.67%

C. Break down denials by payer and by provider

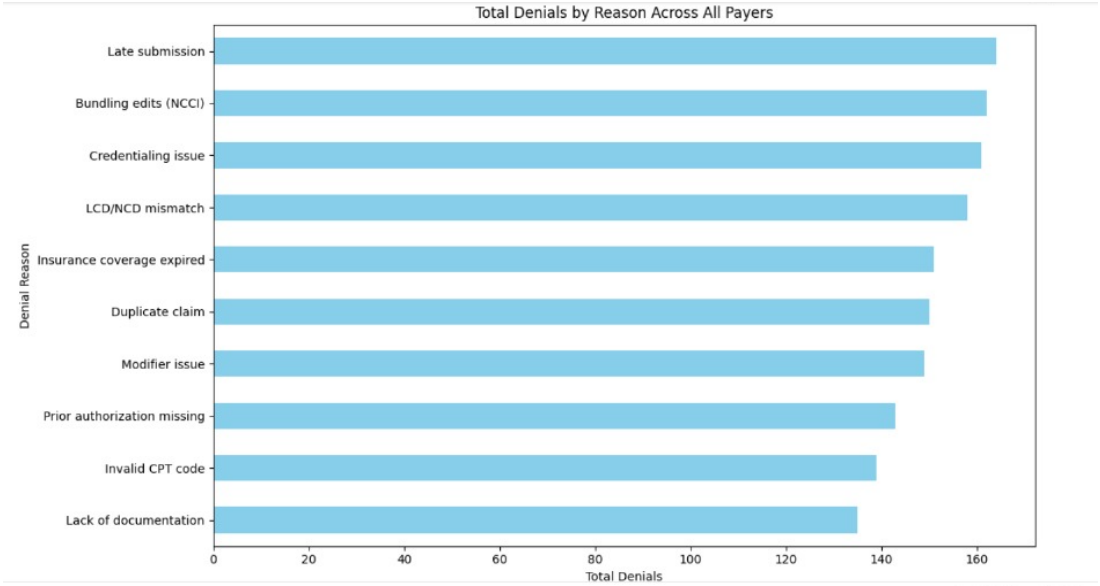
Denials by Insurance Company

Insurance Company	Denied Claims
Humana	165
Aetna	164
Cigna	162
Molina Healthcare	156
Medicare	155
UnitedHealthcare	152
Blue Cross Blue Shield	150
Centene Corporation	149
Anthem	141
Kaiser Permanente	118

Physician	Denied Claims
Dustin Thompson	62
Miss Kristen White	60
Michael Ingram	60
Laura Suarez	59
Karen Hayes MD	55
Amy West	54
Anthony Huynh	53
Jessica Leon	53
Jodi Rodriguez	53
Samuel Jimenez	53
Vincent Gibson	53
Deborah Barker	52
Erin Mclaughlin	52
Gregory Larson	52
Henry Mcneil	52
Caroline Hodges	51
John Young	48
Dakota Black	46
Alyssa Martinez	43
Dr. Jose Richards	43
Lynn Smith	43
Christina Perez	42
Crystal Thomas	42

D. Denial Reasons and Their Frequencies

Denial Reason	Frequency
Late submission	164
Bundling edits (NCCI)	162
Credentialing issue	161
LCD/NCD mismatch	158
Insurance coverage expired	151
Duplicate claim	150
Modifier issue	149
Prior authorization missing	143
Invalid CPT code	139
Lack of documentation	135



Denial Reason Insurance Company	Bundling edits (NCCI)	Credentialing issue	Duplicate claim	Insurance coverage expired	Invalid CPT code	LCD/NCD mismatch	Lack of documentation	Late submission	Modifier issue	Prior authorization missing
Aetna	361	24	17	17	15	16	18	15	10	11	21
Anthem	359	14	12	15	9	9	10	19	17	21	15
Blue Cross Blue Shield	343	17	22	13	16	8	24	9	20	10	11
Centene Corporation	370	17	18	13	20	17	11	16	9	19	9
Cigna	361	12	16	19	14	15	14	15	24	17	16
Humana	319	20	17	22	16	10	18	11	21	15	15
Kaiser Permanente	330	12	9	15	11	15	7	11	14	12	12
Medicare	374	12	13	18	21	14	21	11	9	16	20
Molina Healthcare	334	15	18	8	11	14	17	17	22	20	14
UnitedHealthcare	337	19	19	10	18	21	18	11	18	8	10

2) Detecting the Root Causes

Insurance Coverage Expired – 169 Denials

This appears to be the most common denial. It often means patients didn't have active coverage on the date of service. It could be due to outdated insurance info or gaps in verifying coverage before the visit.

LCD/NCD Mismatch – 168 Denials

These denials suggest that claims aren't aligning with Medicare's coverage policies. It's possible that the diagnosis codes used aren't matching what's allowed under the procedure codes.

Lack of Documentation – 159 Denials

A high number of claims were denied due to missing or insufficient documentation. This could be tied to workflow issues—like providers not uploading supporting notes or the EMR not being properly synced with billing.

Credentialing Issues – 157 Denials

This is usually tied to the provider not being properly enrolled with the payer at the time of service. It can happen when re-credentialing is delayed or not tracked properly.

Duplicate Claims – 155 Denials

These may be caused by billing the same claim more than once—often a result of miscommunication or claim status not being updated quickly in the billing system.

Modifier Issues – 146 Denials

Modifiers may be missing or incorrectly applied. This usually indicates a need for better coding practices or refresher training for the billing team.

Prior Authorization Missing – 146 Denials

These reflect a breakdown in obtaining prior approvals, often due to unclear workflows or last-minute scheduling without insurance checks.

Bundling Edits (NCCI) – 132 Denials

The services in these claims may be bundled under Medicare's NCCI rules. It's important to check which services can or cannot be billed separately.

Recommendations to Reduce Denials:

Review eligibility before every patient visit, even for returning patients.
Ensure the coding team checks LCD/NCD guidelines while billing Medicare claims.
Improve documentation capture and audit records before submission.
Track provider credentialing deadlines and maintain regular updates with payers.
Build in logic to catch duplicate claims before submission.
Conduct training sessions focused on common modifier errors and bundling rules.

3) Recommendations to fix and strategies

Modifier Issues

- Modifiers like -25 and -59 are frequently missed or misused during charge entry.
- Some procedures require specific modifier combinations for correct billing.
- Inconsistent modifier use leads to denials, especially with payers that follow strict edit rules.

Lack of Documentation

- Claims often miss key clinical details that justify the service rendered.
- Incomplete chart notes can lead payers to question medical necessity.
- Documentation gaps are common in high-volume or rushed clinical environments.

LCD/NCD Mismatch

- Services billed don’t align with Medicare’s coverage policies based on diagnosis codes.
- Payers reject claims when necessary codes or documentation aren't included.
- These denials usually affect labs, imaging, and chronic condition procedures.

Prior Authorization Missing

- Services requiring pre-approval are rendered before auth is obtained.
- Authorization numbers are not always linked to the claim correctly.
- Breakdown in communication between scheduling and billing staff can cause delays.

Credentialing or Enrollment Issues

- Claims are submitted under providers not enrolled with specific payers.
- Lapsed or expired credentials can lead to retroactive denials.
- Enrollment delays during new hire onboarding often cause claim holds.

Duplicate Claims & Late Submissions

- Claims may be unintentionally submitted more than once without corrections.
- Lack of tracking systems results in missed payer deadlines.
- Billing teams sometimes confuse rejections with denials and resubmit too early.

Data Summary

- Data Source: Synthetic medical billing dataset created to mimic real-world claim patterns.
- Total Records: ~5000 medical claims
- Key Features Used:
- CPT Code: Procedure codes (categorical)
- Insurance Company: Payer name (categorical)
- Physician: Rendering provider (categorical)
- Amount: Billed amount (numerical)
- Payment: Actual received payment
- Denial Reason (optional): For root cause analysis

Target Variable:

- Claim Status (Denied or Not Denied), derived based on Payment = 0

Preprocessing Steps:

- Handled missing values (if any)
- Encoded categorical variables (using Frequency Encoding and One-Hot Encoding)
- Normalized numerical values (optional depending on model)
- Split data into training and test sets

Data Preparation

• Handling Missing Values:

The dataset was examined for any null or missing values across all columns. Basic data cleaning steps included:

- Dropping rows with critical missing fields such as CPT Code, Insurance Company, or Payment status.
- For less critical fields, imputation techniques (like mode imputation for categorical columns) were considered.

• Feature Encoding:

Categorical variables such as Insurance Company and Physician were encoded using Frequency Encoding. This approach replaces each category with the frequency of its occurrence in the dataset. It was chosen because:

- It preserves information about how common each category is.
- It avoids the dimensionality explosion that comes with One-Hot Encoding.
- It integrates well with tree-based models like Random Forests.

• Feature Selection:

The following features were selected for model training:

- CPT Code: Procedure code indicating the type of medical service.
- Insurance Company: Payer responsible for the claim.
- Physician: The provider who performed the service.
- Amount: The billed amount for the claim.

- Data Splitting:

The cleaned and encoded data was split into training and testing sets:

- Training Set (80%): Used to train and cross-validate the model.
- Testing Set (20%): Held out to evaluate the final model performance and generalizability

Model Development

• Algorithm Used:

The primary machine learning algorithm employed was the Random Forest Classifier, chosen for its robustness, ability to handle mixed data types, and effectiveness with tabular data. Random Forests are ensemble models that reduce overfitting by averaging predictions from multiple decision trees.

• Feature Selection and Engineering:

The model was trained using the following features:

- CPT Code (procedural code)
- Insurance Company
- Physician
- Amount

Categorical features like Insurance Company and Physician were encoded using frequency encoding, which maps each category to the frequency of its occurrence. This is effective in models like Random Forest that can handle numerical representations of categorical data without assuming ordinal relationships.

• Hyperparameter Tuning:

To optimize model performance, GridSearchCV was used with 5-fold cross-validation. Parameter such as:

- n_estimators (number of trees),
- max_depth,
- min_samples_split, and
- min_samples_leaf

were tuned to find the combination that yielded the best validation results.

• Performance Metrics:

- Best Accuracy (Cross-Validation): 75.5%
- Test Accuracy: 79.2%
- Precision & Recall: Balanced scores indicating the model is effective at correctly identifying both denied and non-denied claims.
- F1-Score: Maintained around 0.77–0.81 across classes, suggesting good overall model balance.

Conclusion:

The Random Forest model with tuned hyperparameters achieved reliable and generalizable performance for predicting claim denials, making it a suitable choice for deployment in real-world RCM workflows.

6. Application Feature: Batch Prediction

Bulk Upload:

- The application allows users to upload Excel or CSV files containing multiple medical claims at once. This enables efficient processing of large datasets without manual entry for each claim.

Automated Processing:

Once the file is uploaded, the system automatically:

- Validates the presence of required columns such as CPT Code, Insurance Company, Physician, and Amount.
- Applies the pre-trained frequency encoder to transform categorical data.
- Uses the trained machine learning model to predict claim denial status for each record.

Downloadable Output:

After prediction, users can download the results as a CSV file. This file includes the original claim data along with an added column indicating whether each claim is predicted as "Denied" or "Not Denied."
This feature supports easy integration with existing workflows and enables further analysis or reporting

Tools & Technologies

- Programming Language: Python
- Libraries: pandas, scikit-learn, Streamlit, Mattlotliv
- Model & Encoder Storage: pickle
- User Interface: Streamlit Web Application