MedEvalPro

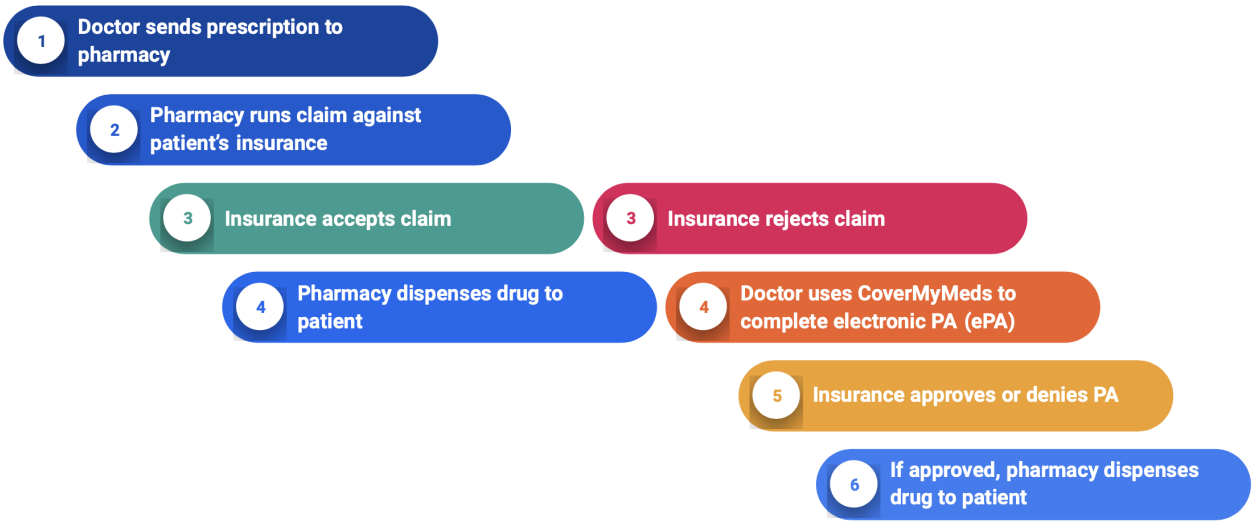
Medical Evaluation Prophet: a Forecasting Tool for Medical Claims

The Erdös Institute Data Science Bootcamp, Fall 2022

Ismail Abouamal, Po-Wen Chang, Özkan Demir, Axel La Salle, and Victoria Uribe

1 Overview

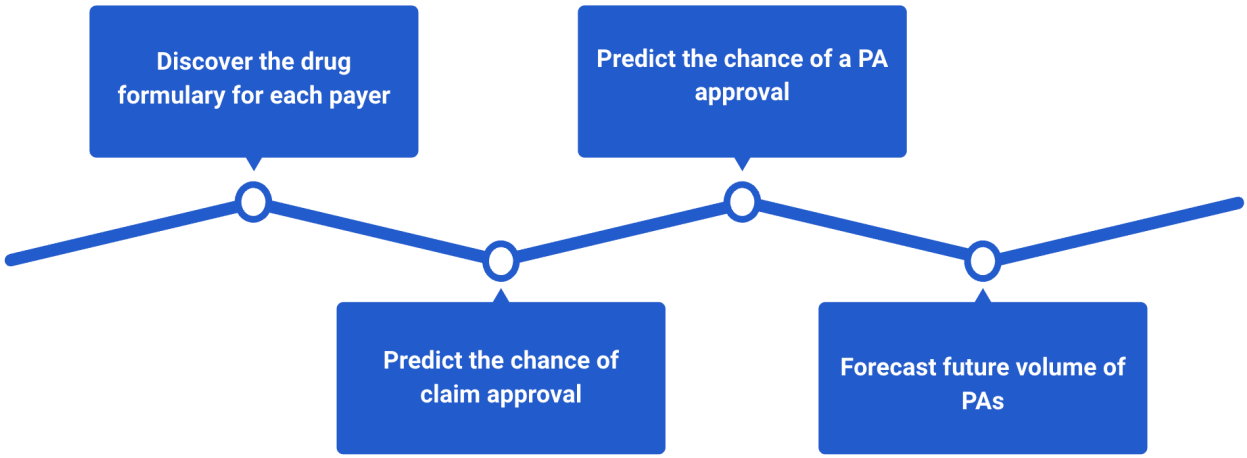
CoverMyMeds is a healthcare technology company that aims to improve patients' access to medications by simplifying the process of connecting with the healthcare network. It can speed up the time for patients to receive therapy and reduce the number of prescriptions that will not be approved by the health insurance payers. Furthermore, it provides the electronic prior authorization (ePA) service for the medical claims rejected by payers, which could tremendously boost the process of PA filing with a convenient portal-based experience.



Taking advantage of the historical medical data associated with patients' claims and PAs, it is possible to extract valuable information and make practical applications to benefit customers. For

example, we can infer payers' formulary of drugs and how payers make decisions based on it and medical prescriptions.

Through this project, we provide Medical Evaluation Prophet (MedEvalPro) , a forecasting tool that enables healthcare providers to make efficient decisions on whether or not a PA is needed, significantly enhance the chance of PA approval, and predict the trend of future demands on PAs. We use the data provided by CoverMyMeds to perform the following tasks:



2 Project Summary

In this section, we summarize the main results we find. We first discover the drug formulary of each payer. This will help us gain some important insights for our dataset. We then train classification models, predict the chance of claim approval and PA approval, compare the models, and highlight our findings. We further show our forecast of the future PA volumes using the time series analysis.

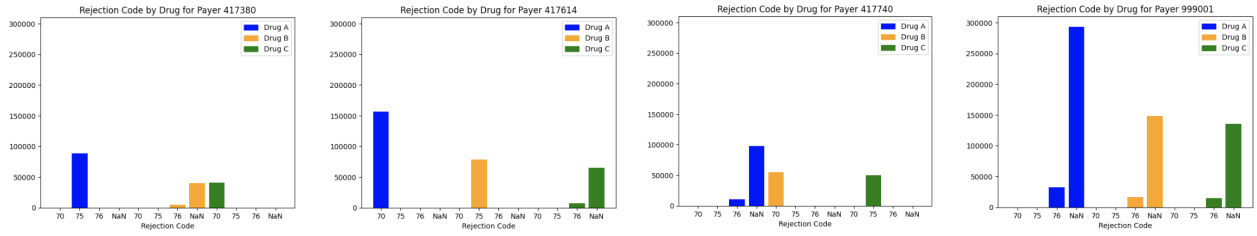
2.1 Formulary

Claims with specific drugs could be rejected for various reasons. In our data, there are three reject codes:

- A code 70 means that the drug is not covered by the plan and is not on the payer's formulary.
- A code 75 means that the drug is on the payer's formulary but is not preferred and a PA is needed.
- A code 76 means that the drug is covered but that the plan's limit on the number of fills for that drug have been met.

We are able to discover the formulary of each payer. We find all drugs (A, B, C) are associated with different reject codes (70, 75, 76) by bin 417380, 417614, 417740, while bin 999001 rejects all

drugs by code 76. This implies that the plans of bin 999001 cover all three types of drugs, but may have relative low limit on the number of fills for drugs.



Drug \ BIN	417380	417614	417740	999001
A	75	70	76	76
B	76	75	70	76
C	70	76	75	76

Figure 1: A table for each health insurance payer's (labeled by "bin") formulary. The drugs associated with code 70 or 75 (dark blue blocks) will never be approved at the level of pharmacy claim.

2.2 Classification

Our classification tasks are two-fold:

1. To classify what kinds of claims are likely to be approved or rejected.

2. To predict what kinds of PAs will be approved or rejected.

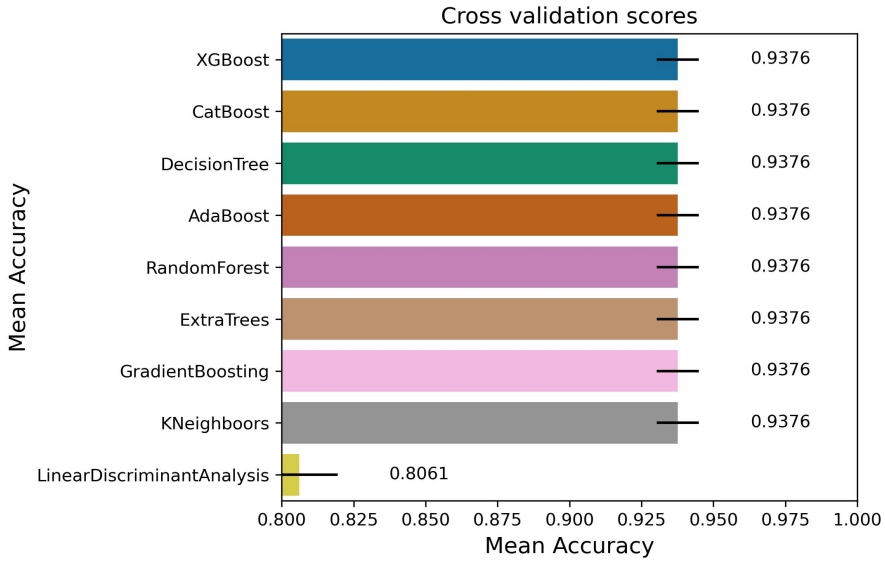
With our data, we train the following classification models: Decision Tree, AdaBoost, Random Forest, Extra Trees, Gradient Boosting, K Neighbors, Logistic Regression, Linear Discriminant, CatBoost, XGBoost, feedforward neural network (consists of two layers of 50 and 20 nodes respectively and uses the rectifier function as the activation function).

Classifying claims

Can we predict whether a claim will be directly approved by payers? If the answer is positive, the provider will know if a PA is needed when making a prescription, which could save time for patients.

As there are only 2 relevant features ("bin_no" and "drug") and the final label (i.e., "pharmacy_claim_approval") may be affected by many factors not included in the dataset, it could be difficult to make a robust classification.

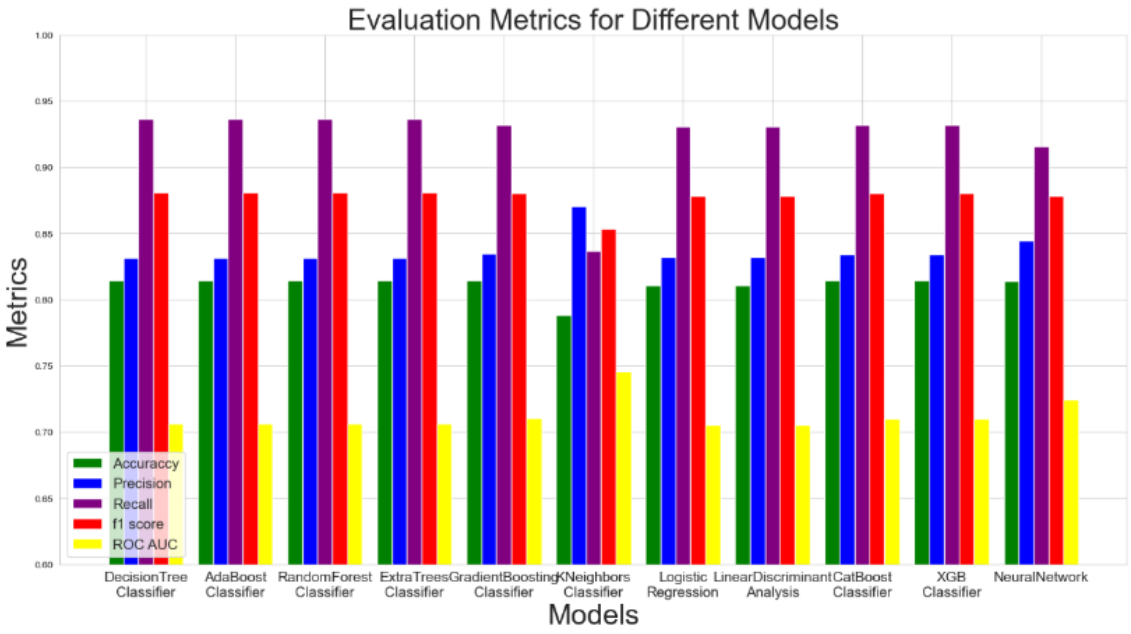
Surprisingly, it turns out that we can actually do an excellent job on this task! We find that most classification models can achieve 94% of accuracy in predicting the claim approval. We find that, this



is because a claim will *never* be approved if its drug is not on a payer's plan or is not preferred (i.e., those drugs associated with reject code 70 or 75 in Fig. 1). We conclude that "bin_no" and "drug" are robust predictors for the chance of claim approval!

Classifying PAs: The metric used for the comparison of all models

With all rejected data, we can also predict what kind of PAs will be approved or rejected. We have the results comparing above models on a single train-test split on the whole dataset. The input for our classification models is an 8-dimensional array containing the categorical entries corresponding to the following features: "rejected_code", "drug_type", "correct_diagnosis", "tried_and_failed", "contraindication", "bin_number."

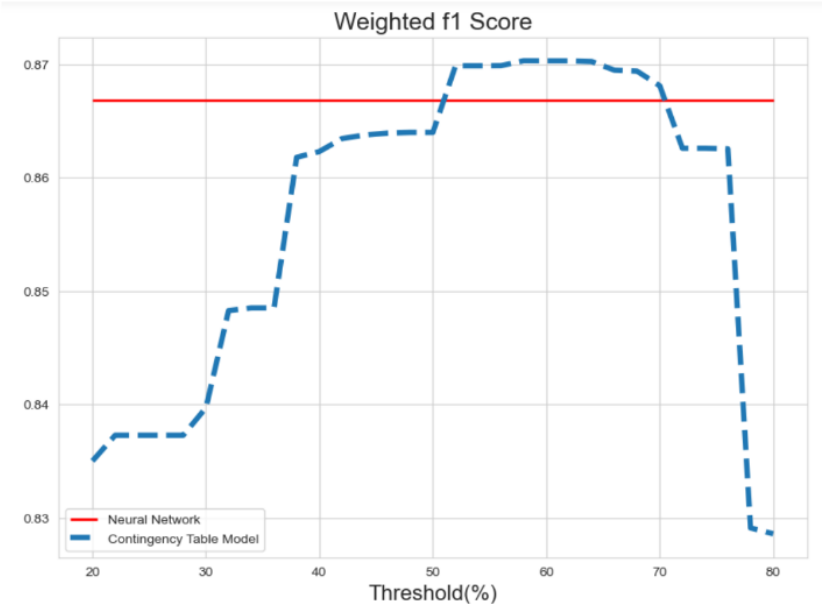


In summary, for our metric to choose the final model, we did not use accuracy because the dataset

was imbalanced, having 73% PA approval rate overall. For this classification problem, false positive is worse than a false negative, as classifying a to-be-rejected PA claim as otherwise and filing the PA claim would put an unnecessary cost on the company. So, precision is our most important metric to maximize. However, the naive model that rejects every PA claim has 100% precision, so we need another metric. We decided to use a weighted harmonic mean of precision and recall, weighed more on the precision side (2-to-1), which can be thought of as a weighted f1 score. With this metric, the Feedforward neural network model turned out to have the best performance.

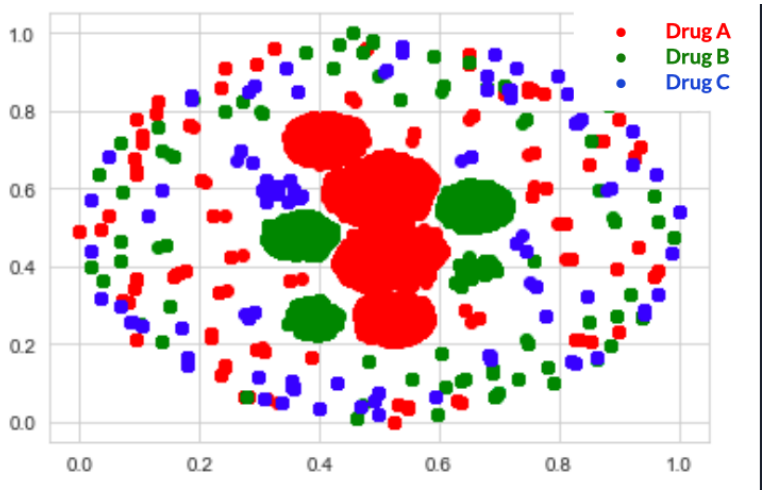
Classifying PAs: an intuitive model using PA approval rates calculated from contingency tables

In addition to these machine learning models, we computed PA approval rates using contingency tables and build a model by using different thresholds with the approval rates determined by our training data. We observed that using all six categorical features we used in the above models results in a contingency table with cells with 0 counts, and those cells cannot be used to compute approval rates. So, we used the feature importance scores we get from the best model to get an importance order among the subsets of features, and used it to get a contingency table with a nonempty count for any given feature string. The following are the top 4 feature subsets in this order: All features > All features except "correct_diagnosis" > All features except "bin_no" > All features except "correct_diagnosis" & "bin_no" We observe that the contingency table built from all features except "correct_diagnosis" & "bin_no" is full, so we only needed to go this deep in the order to get all the approval rates. Then we get the weighted f1 scores for models with different thresholds, and show that the model with 60% threshold outperforms all other machine learning mothers we have investigated so far. This is caused by the fully categorical nature of the dataset.



Classifying PAs: Embedding of the last layer of the neural network

Is our best model capable of separating the claims based on a given feature? We used the TSNE (t-distributed stochastic neighbor embedding) algorithm to reduce the dimension of the last layer of the network, which is a 20-dimensional array. This allows us to visualize how the model performs the classification task. For example, based on the figure above our feedforward network seems to isolate the claims containing drugs A and B relatively well.

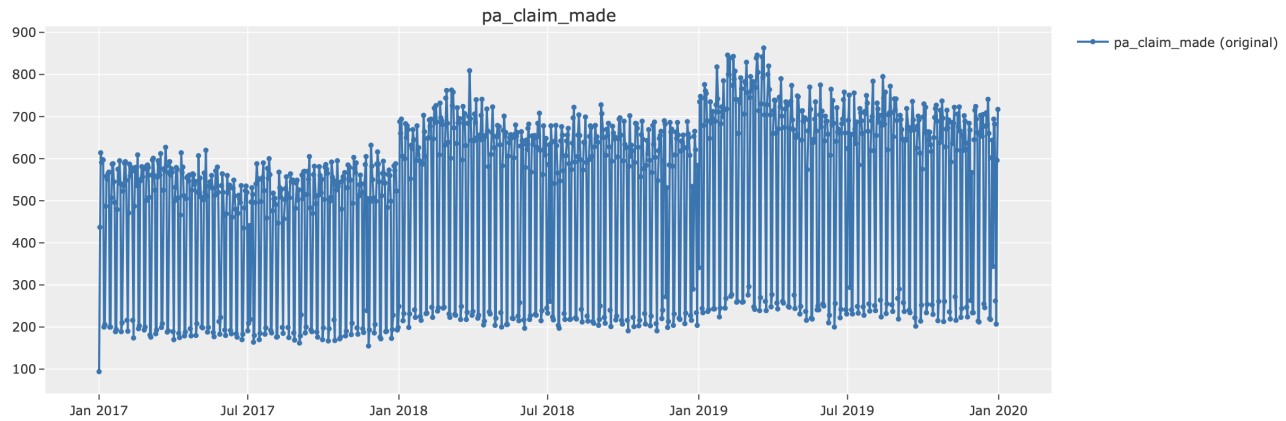


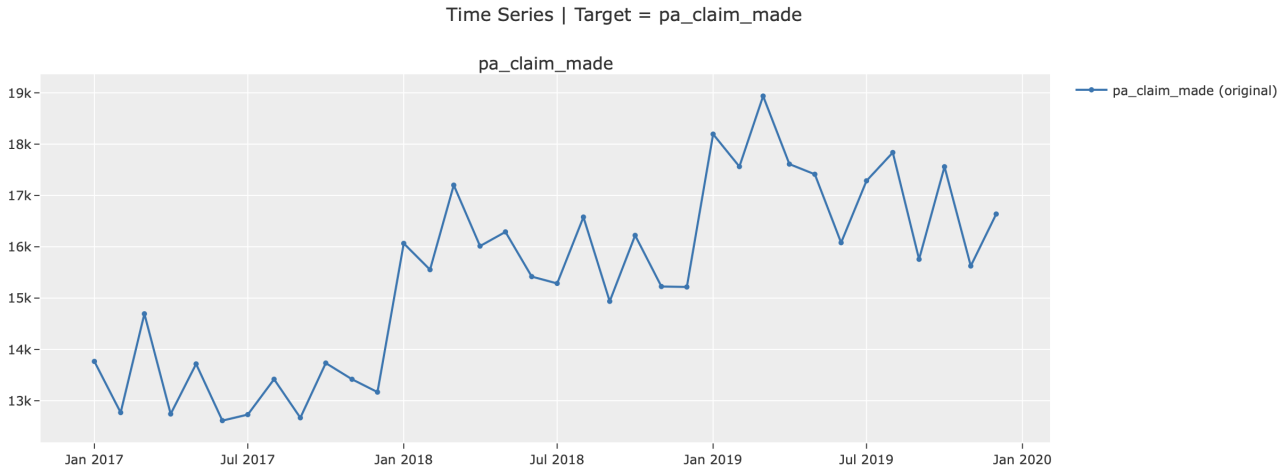
2.3 Time Series Analysis

Motivation

If a claim is rejected, then regardless of the rejection code, an ePA can be started to ensure the patient recieves the care they need. At CoverMyMeds, an automated system is used to generate ePAs. From a financial perspective, ePA volume is a predictor of revenue. We forecast future volume (at the daily/monthly/yearly level) using time series analysis to help with budgeting. We often see ePA volume at its highest during workdays, and its lowest during holiday seasons. We also see peaks in volume around the start of the year, which is when many PAs expire and new ones will be resubmitted.

Data





Key question

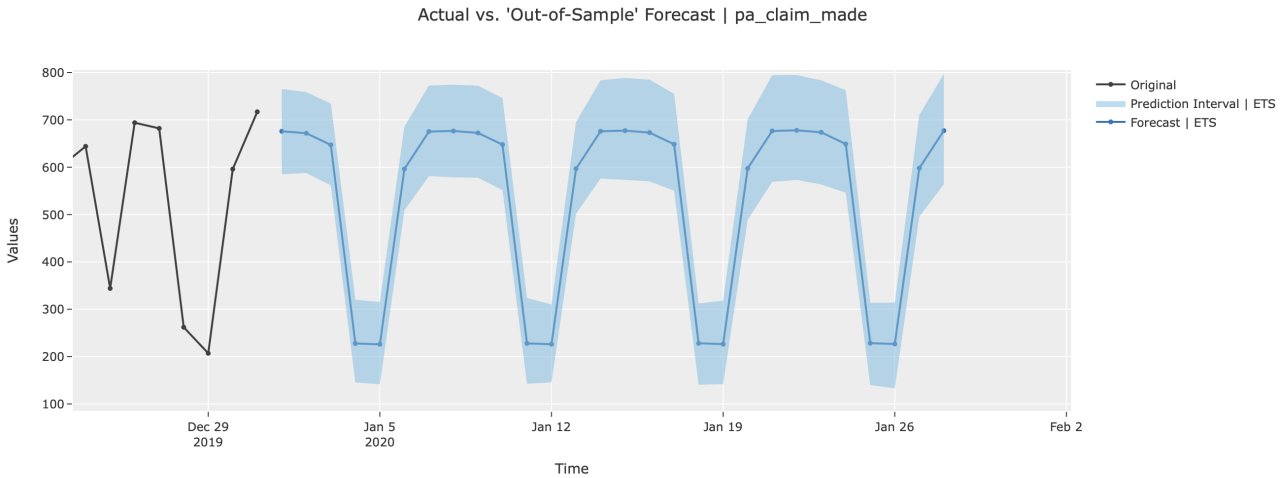
How to forecast ePA volumes in the future? We seek to find the best models for future forecasting given historical ePA volume data.

Best models

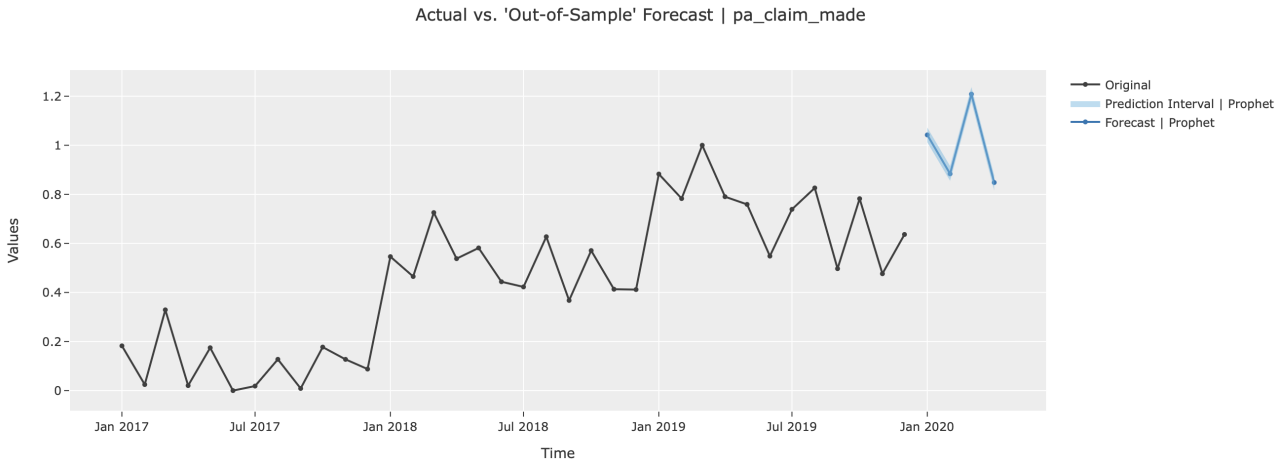
- For daily forecasting - ETS.
- For monthly forecasting - Prophet.
- Using exogenous features - Auto ARIMA

Results

- Future forecast - daily level



- Future forecast - monthly level



Performance Summary

- ETS - 0.0678 (Mean Absolute Percentage Error).
- Prophet - 0.1104 (Mean Absolute Percentage Error).
- AutoRegressive LSTM - 0.1991 (Mean Absolute Error).