# Models Performance Report


*Ahmad Alqaisi*

Sunday, July 21, 2024

# Outline

- Recap
- Current Model’s Performance
- Data Pre-processing
- ML Model Performance
- Next Step
- Recommendation

# Recap

Collection performance, in general, over the past years has been noticeably low.

As only 11% of the total collections assigned to each collector are collected from the claims values per claim, which indicates

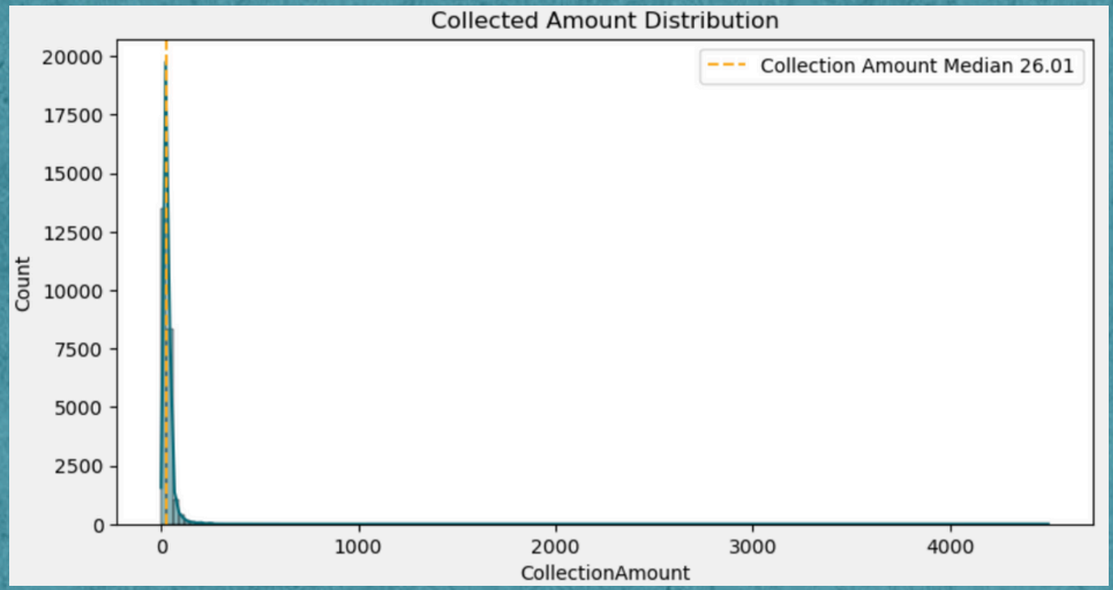
Claims values are also declining, which consequently affects the collections values in a linear fashion.

Well, the collection values are also related to the number of connections to the debtors quadratically.

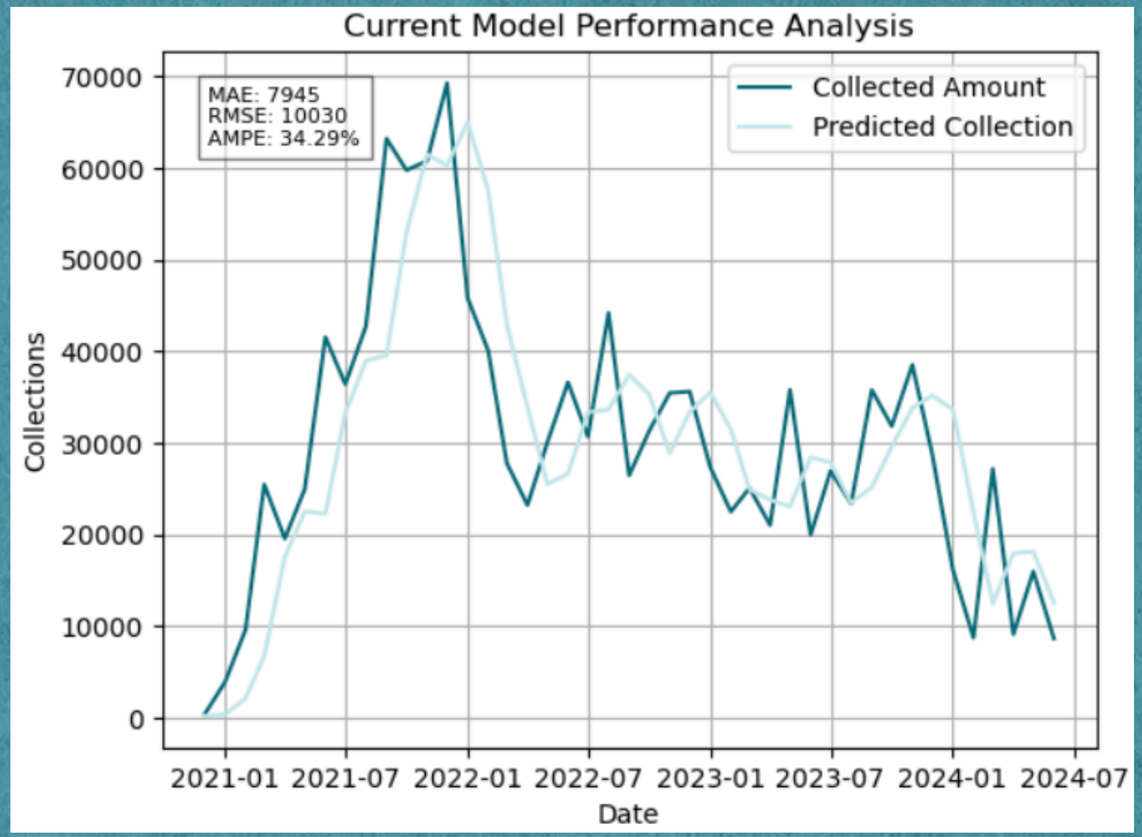
# Current Model's Performance

The model currently used predicts collection values by taking the average of collections for the past two years.

In fact, it may seem a little feasible, but the average is sensitive to outliers, and as we saw previously, the distribution of claims values was skewed to the right, and there are also a fair number of outliers, which will affect the result of calculating the average and make it inaccurate.



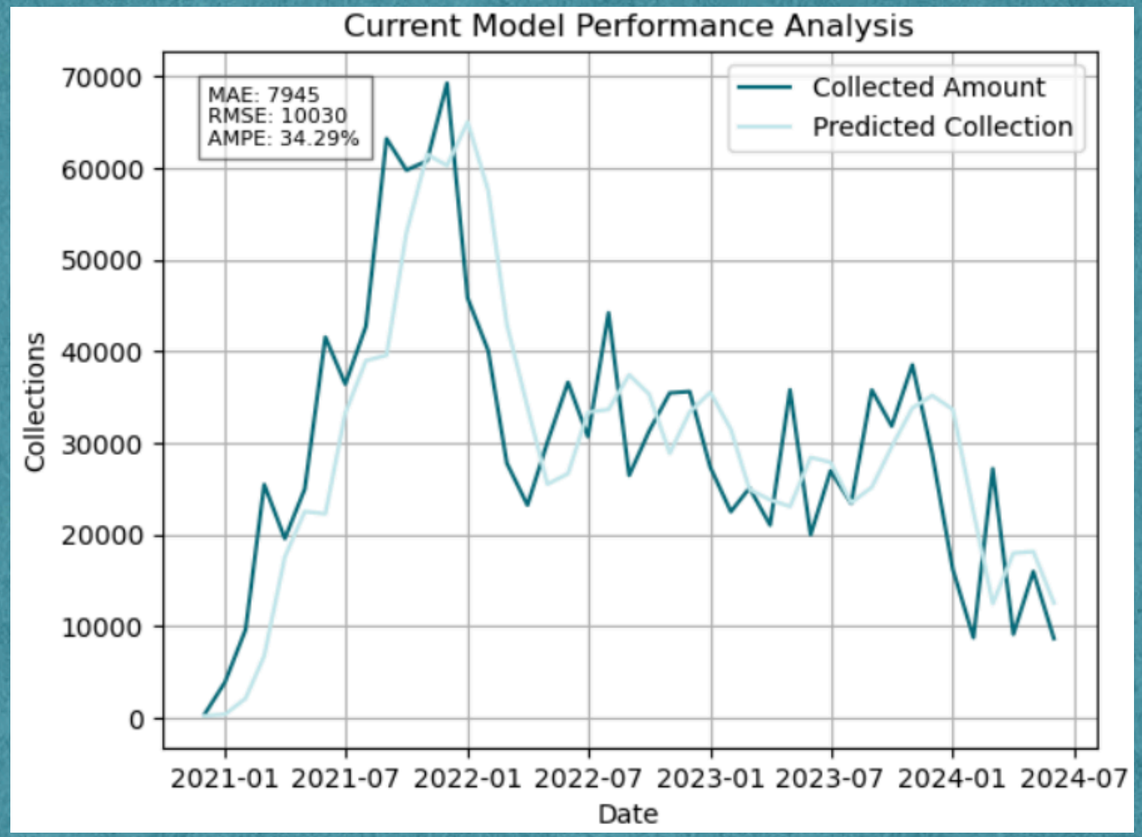
# Current Model's Performance




The model roughly follows the line of collection. It is obvious to notice this behavior of the model since it is based on just finding the average.

However, it has a value of mean absolute error (**MAE**) about 8000, which means on average the predictions are off by 8000 from the actual collection values.

# Current Model’s Performance




A Mean Absolute Percentage Error (**MAPE**) of 34% indicates that, on average, the model's predictions deviate from the actual values by 34%.

Lastly, a model has a value of root mean squared error (**RMSE**) about 10000, reflecting larger errors more significantly.

# Data Pre-processing

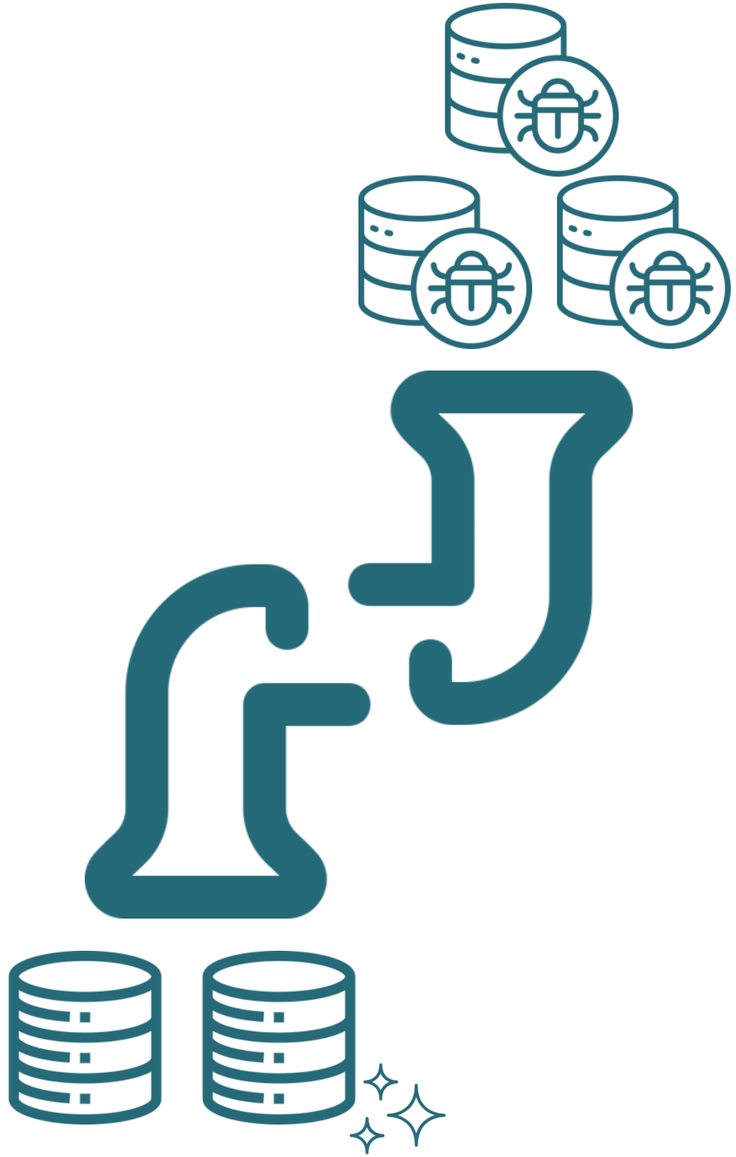
The next step after exploring and visualize the data is to prepare it for a machine learning modeling.

This stage involves deleting or substituting artificial values in place of the missing values. I used the nearest neighbor model to do this process.

Due to the lack of available data related to predicting collection values at forecasting time, such as the number of connections, I did *feature engineering*.

I calculated new metrics in tables based on historical collection values, such as how the conductor contributes to the total daily collection values, what is the average number of calls the conductor makes daily, and others to help in the forecasting process.

Please note that these tables must be updated frequently to maintain the model's performance over time

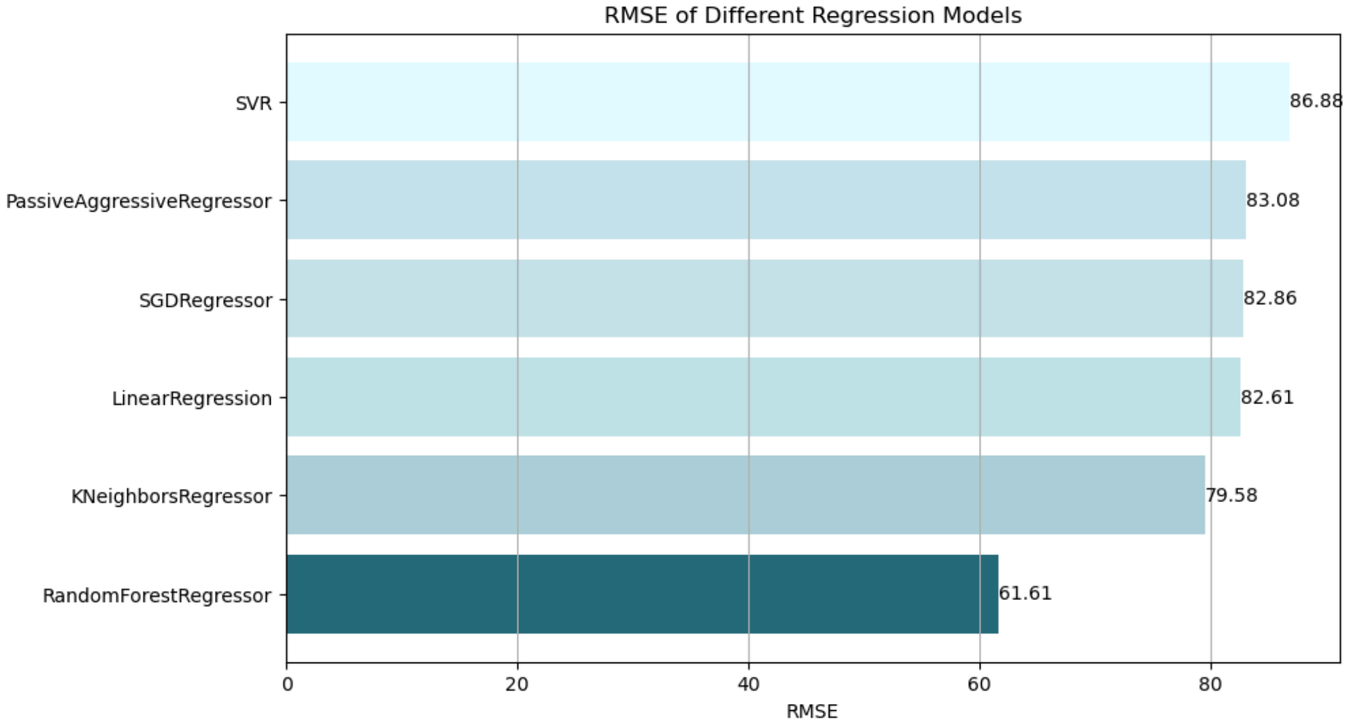


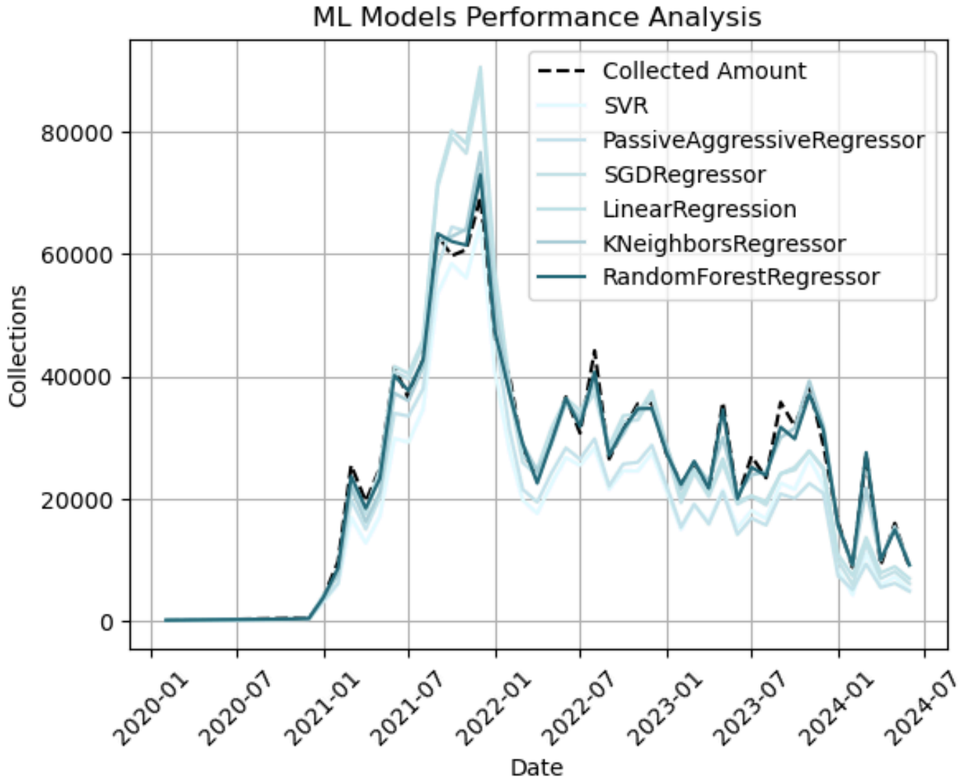
# ML Models Performance

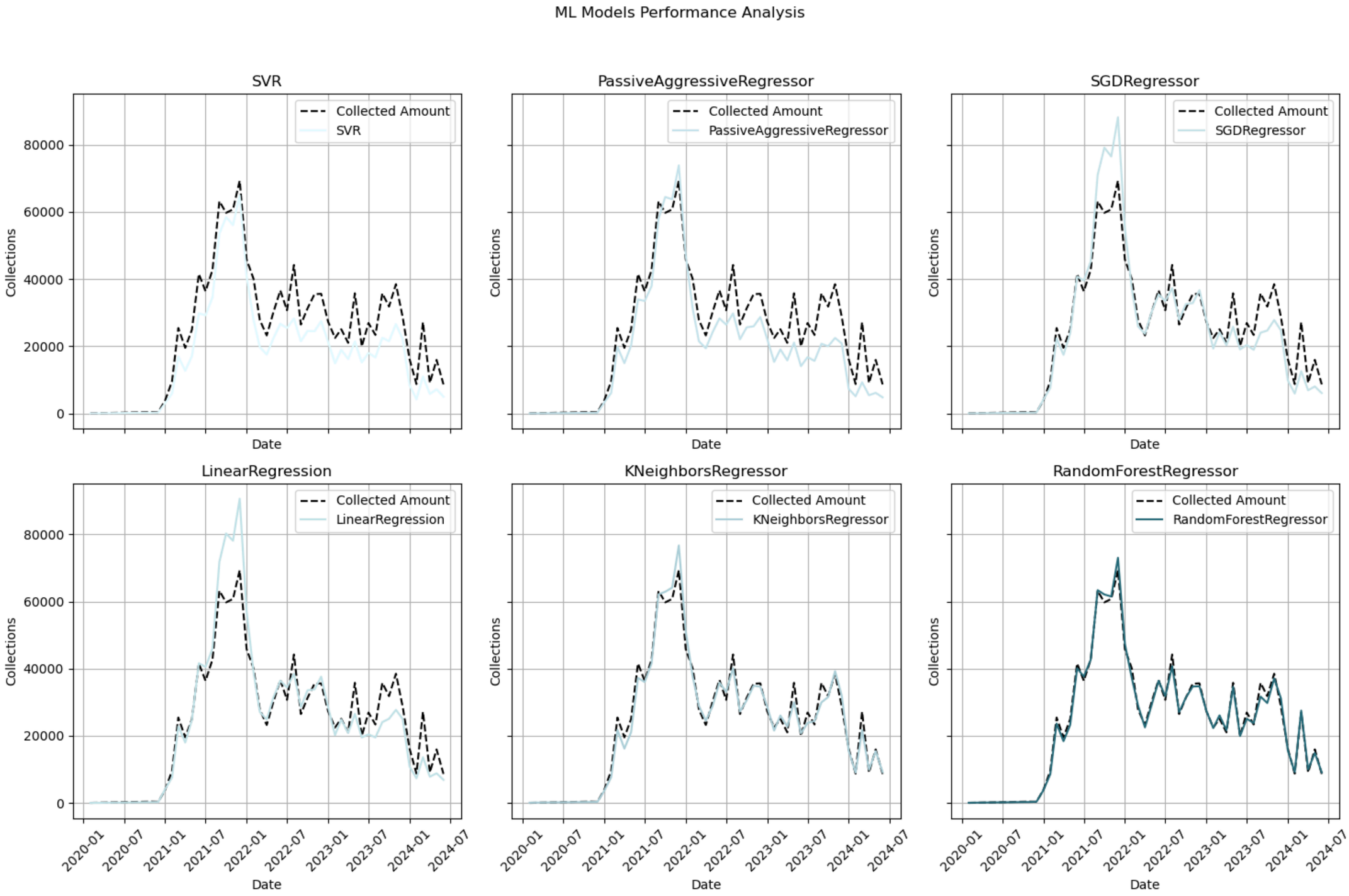
The goal of this result is to come up with up to 5 models then combine them in one great solution.

The below table list some models with their corresponding cost score values:

| | RMSE | AME | AMPE |
|---|---|---|---|
| **SVR** | 86.88 | 18.97 | 2.59% |
| **KNeighborsRegressor** | 79.58 | 18.67 | 2.88% |
| **PassiveAggressiveRegressor** | 89.03 | 20.81 | 2.92% |
| **RandomForestRegressor** | 62.75 | 15.48 | 3.23% |
| **LinearRegression** | 82.61 | 22.35 | 3.33% |
| **SGDRegressor** | 83.06 | 22.58 | 3.40% |



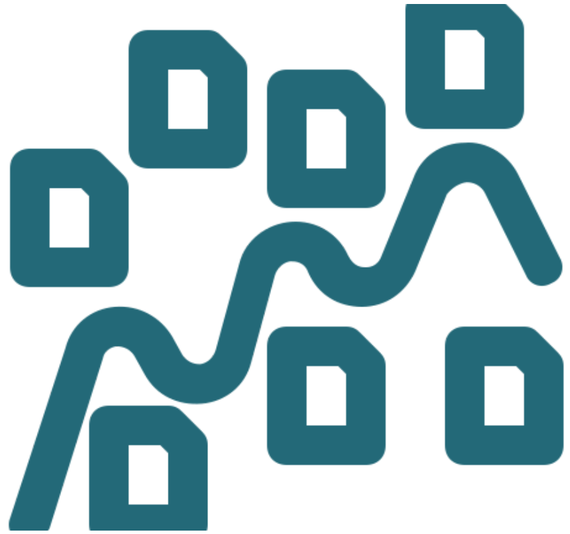




ML Models Performance Analysis
SVR
PassiveAggressiveRegressor
SGDRegressor
LinearRegression
KNeighborsRegressor
RandomForestRegressor
Collected Amount
SVR
PassiveAggressiveRegressor
SGDRegressor
LinearRegression
KNeighborsRegressor
RandomForestRegressor
Collections
Date
0
20000
40000
60000
80000
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
2024-07

# Professionalism is the enemy of goodness

One more secure approach is simply to model the relationship between the amounts and the collected amounts instead of modeling based on previously new calculated complex metrics based on the historical claims, which may *overfits* the trained models.

| | RMSE | AME | AMPE |
|---|---|---|---|
| **PassiveAggressiveRegressor** | 111.66 | 35.78 | 1.30% |
| **SVR** | 85.43 | 19.64 | 2.13% |
| **RandomForestRegressor** | 61.17 | 19.92 | 2.18% |
| **KNeighborsRegressor** | 70.21 | 20.11 | 2.25% |
| **LinearRegression** | 82.70 | 21.84 | 2.76% |
| **SGDRegressor** | 3764598500260807.00 | 1289213664730106.25 | 82817982906155.73% |



The stochastic gradient descent (**SGD**) model is extremely *underfits* the train dataset as the same in the other liner models which have the highest **RMSE** values.

# Next Step

The next step is to fine tune the trained models and integrate them in one ensemble mode.

Well, we can skip this phase, If we adopt the first approach, since we have reached reasonable scores and we can start to deploy the model to be used. But if we adopt the second approach, then we should take this phase into account.

# Recommendation

Congratulations, we have achieved the business case goal of building a model that outperforms the current model.

But which approach we should adapt ?
Well, it’s depends. The first approach is based on calculating new metrics from the historical claims data which overfits them which affects the predictions in the future and for this reason these tables should be updated frequently to the pipeline to maintain the model performance.

# Recommendation

On the other hand, if we adapt the second approach, it lacks context. This method is simple enough so it doesn't overfits the training dataset and it’s only models the current relationship between claims and collections based on historical performance.

For example, if the average collectors' contribution was about 11% of the total claim values, and this contribution increases, the model will not adapt to these changes. It will continue making predictions based on the previous collectors' performance (11%) it found in the historical data ending up with underfits the training set.

# Recommendation

We can run both approaches in parallel.

We can use the first approach and update the tables periodically or we can find new metric more relevant to the prediction may by conducting a data analysis projects.

# Thank you!


*Ahmad Alqaisi*

Sunday, July 21, 2024

# Contact


0785204711

amq4319@gmail.com

https://www.linkedin.com/in/ahmad-al-qaisi/