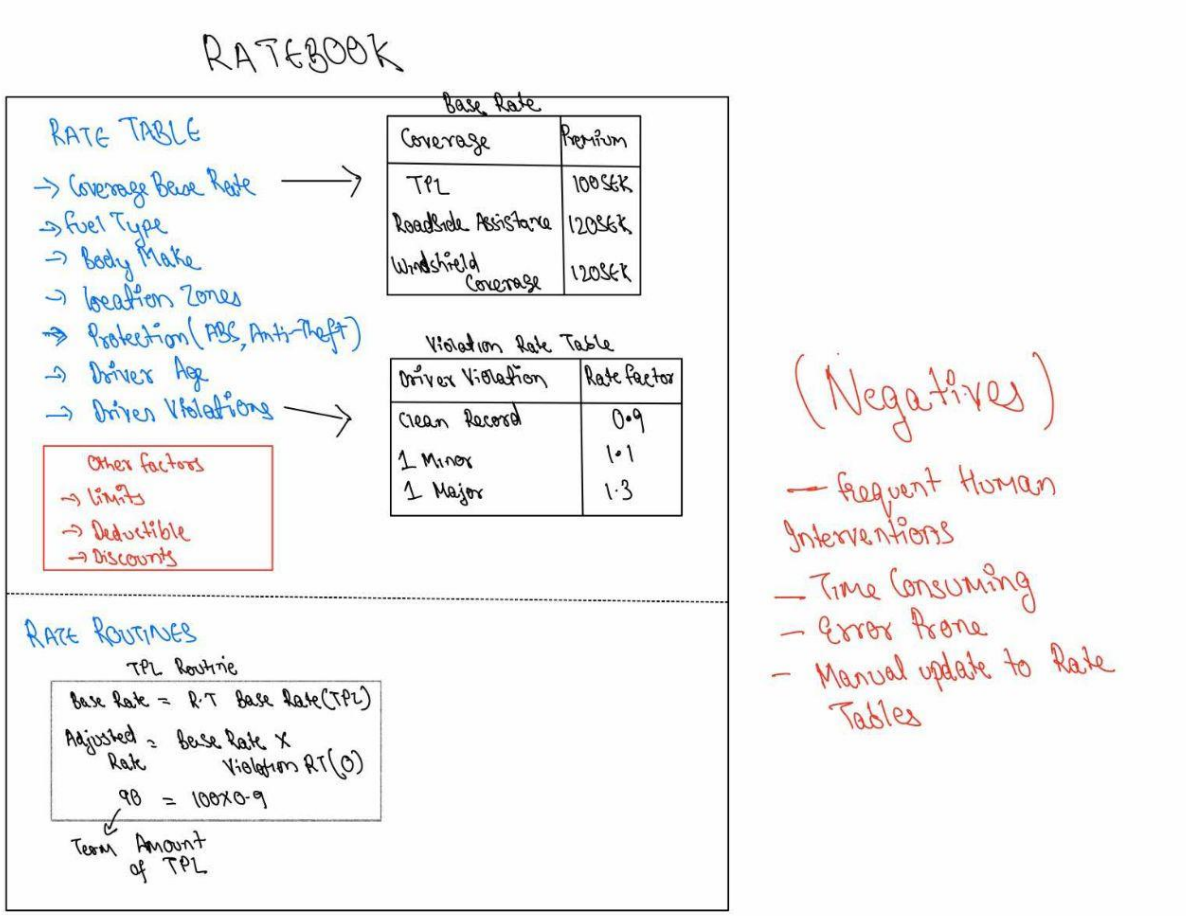
INTELLIGENT CLAIM FREQUENCY MODELING

Data-Driven Insights for Improved Pricing

AMAN JAIN
13.01.2025

PROBLEM STATEMENT

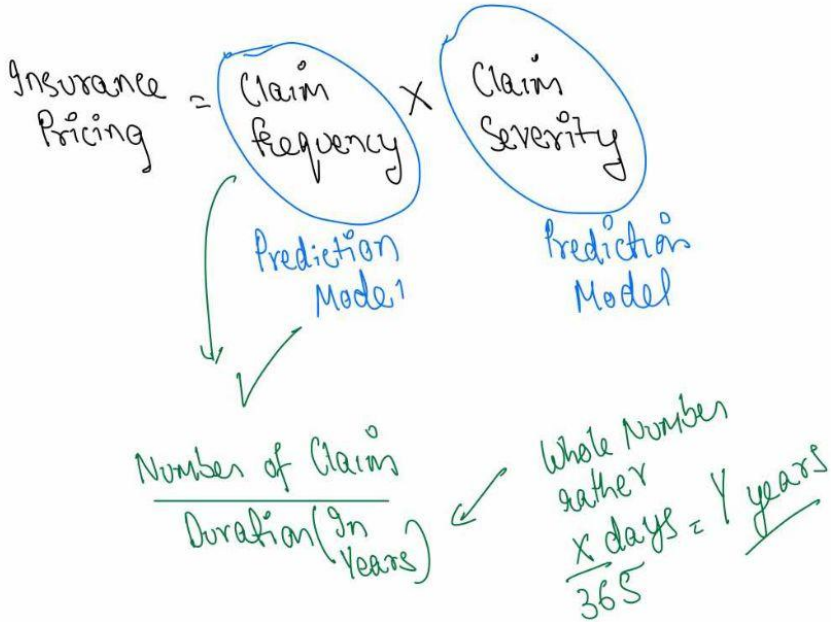


(Negatives)

- Frequent Human Interventions
- Time Consuming
- Error Prone
- Manual update to Rate Tables

SOLUTION

AUTOMATE THE PRICING FRAMEWORK



DATA

- owner_age: the age of the owner
- owner_gender: the gender of the owner, M (male) or K (female).
- geo_zone: geographic zone numbered from 1 to 7, based on the address of the owner
- vehicle_class: a classification based on engine power, vehicle weight, 7 classes in total
- vehicle_age: vehicle age, between 0 and 99
- duration: the number of policy years
- n_claims: the number of claims
- claim_cost: the claim cost

While exporting the data from the system it could be Claim Group, or Paid/Unpaid Claim

n_claims	Claim_cost (Average)
1	18364
1 1	36645
2	39445
2 1 1 2	47000
2 2	76298

Researching About the Data

this to give us good an estimate as you can.

Local crime rates

An increase in vehicle or part theft or damage where you live can affect your premium even if you haven't been directly affected.

If you live in an area where this is happening, take steps to prevent something similar happening to you like parking your car in a safe space or increasing security around your home.

Your insurer will see you as better protected and will be less likely to bump up what they quote you.

It might be worth buying an alarm for your car if it has a catalytic converter, for example.

Vehicle

Your vehicle will determine the estimated cost of claims repair or replacement. Generally, the higher the cost, the higher your rates will be as a result.

Some features such as anti-lock brakes and anti-theft devices may help reduce your rates; however, it's important to note that others, such as advanced safety features with sensors for crash avoidance, might increase your rates since they rely on technology that's often more expensive to repair or replace than their traditional counterparts.

Location

Where you live can influence how often you'll be exposed to potential risks such as accidents, theft and vandalism. The more risks you could be exposed to, the higher your rates may be.

For example, your rates will typically be higher if you live in an area with heavy traffic and potentially more risks vs. an area with less traffic and fewer risks.

Age

Your age is important because it helps companies estimate how many accidents you could potentially have due to the strong correlation between driving age and accidents.

u may have lower rates when you're between the ages of 25 and 65 because drivers in that age range nd to have fewer accidents overall. On the other hand, you might have higher rates when you're under 25 over 65 because of the increased likelihood of having accidents and sustaining injuries.

TOOLS AND TECHNIQUES

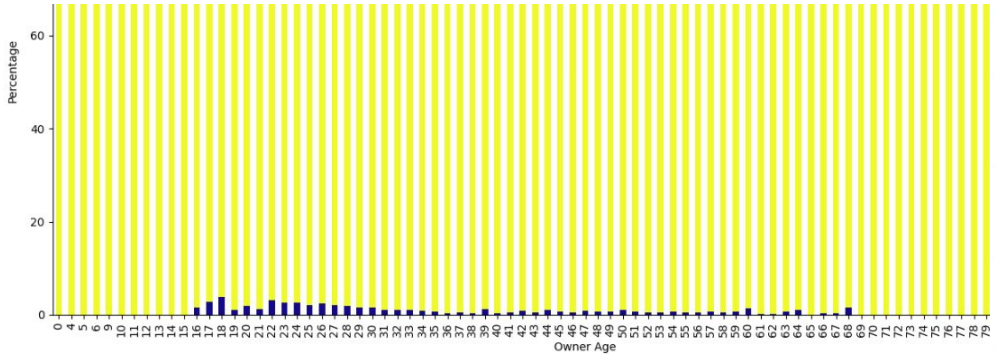
- **Tools**:
  - Figma - for intuitive diagrams
  - Notability - Notes
  - Python (Pandas, Scikit-learn, Matplotlib/Seaborn).
  - Jupyter Notebook (Gemini AI Integrated - For utility scripts)
  - Libraries for Predictions (Linear Regression).
  - ChatGPT - for redirected towards the data/research papers
  - ResearchGate - For research papers
  - Google Slides - for presentation
  - Google Sheets - for exploring the datasets

- **Techniques**:
  - Supervised Regression Technique
  - Linear Regression Model

EXPLORING THE DATA

claim_made	No	Yes	Total	Yes%	No%
geo_zone					
1	8293	173	8466	2.043468	97.956532
2	11425	162	11587	1.398119	98.601881
3	12341	118	12459	0.947107	99.052893
4	24088	190	24278	0.782602	99.217398
5	2338	9	2347	0.383468	99.616532
6	3826	17	3843	0.442363	99.557637
7	369	1	370	0.270270	99.729730

claim_made	No	Yes	Total	Yes%	No%
owner_gender					
0	9701	60	9761	0.614691	99.385309
1	52979	610	53589	1.138293	98.861707

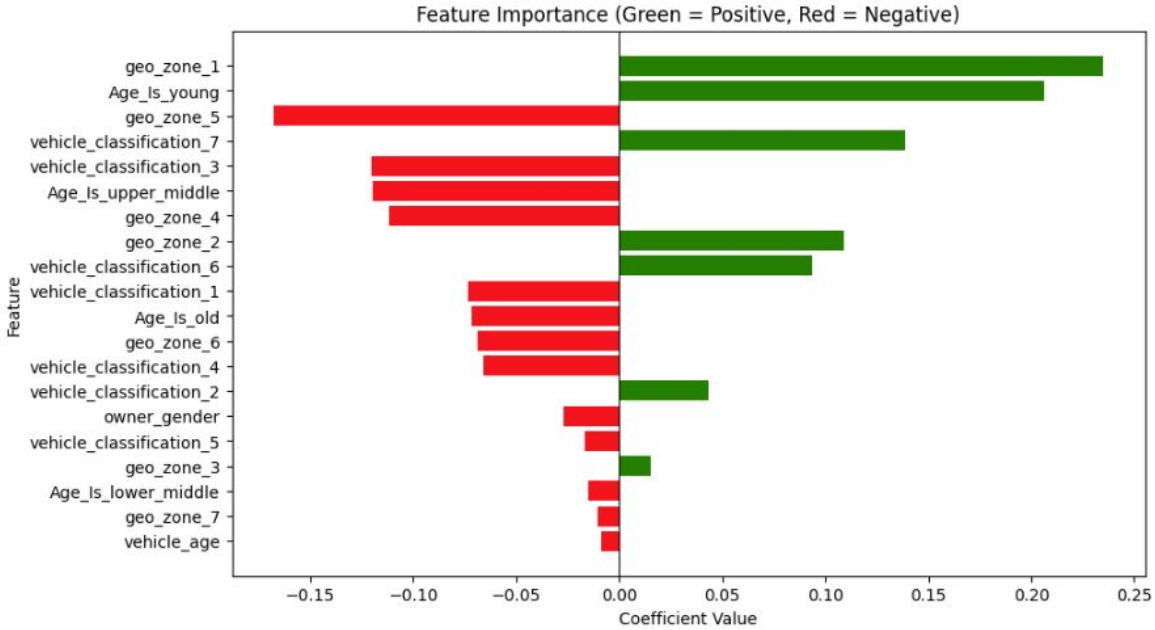


claim_made	No	Yes	Total	Yes%	No%
vehicle_classification					
1	6853	45	6898	0.652363	99.347637
2	5059	57	5116	1.114152	98.885848
3	18295	158	18453	0.856229	99.143771
4	12022	93	12115	0.767643	99.232357
5	11516	144	11660	1.234991	98.765009
6	8165	168	8333	2.016081	97.983919
7	770	5	775	0.645161	99.354839

LINEAR REGRESSION METRICS

Model	MSE	R^2
Linear Regression with All Features	0.22186	0.267313
Linear Regression without Owner Gender and Vehicle Age	0.227047	0.250185

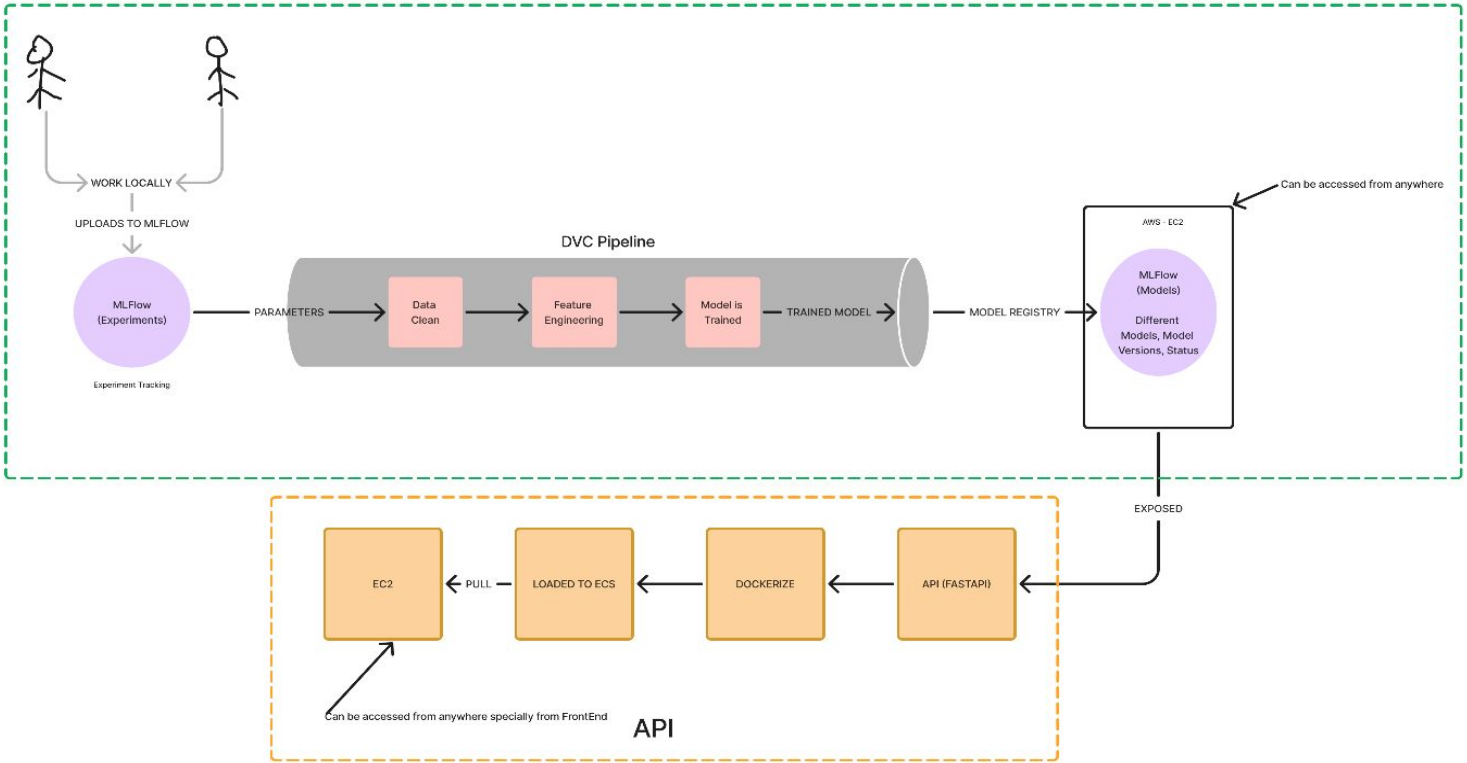
FEATURE IMPORTANCE



CONCLUSION

- **Conclusion**:
  - Summary of key findings.
  - Geo Zone 1 is the riskiest and 5 is the safest
  - Young car owners tend to have more claims
  - With more time, the confidence on the model can be improved by hyper parameter tunings and more feature engineering

NEXT STEPS


WORK LOCALLY
UPLOADS TO MLFLOW
MLFlow
(Experiments)
Experiment Tracking
PARAMETERS
DVC Pipeline
Data
Clean
Feature
Engineering
Model is
Trained
TRAINED MODEL
MODEL REGISTRY
Can be accessed from anywhere
AWS - EC2
MLFlow
(Models)
Different
Models, Model
Versions, Status
EXPOSED
EC2
PULL
LOADED TO ECS
DOCKERIZE
API (FASTAPI)
Can be accessed from anywhere specially from FrontEnd
API

THANK YOU