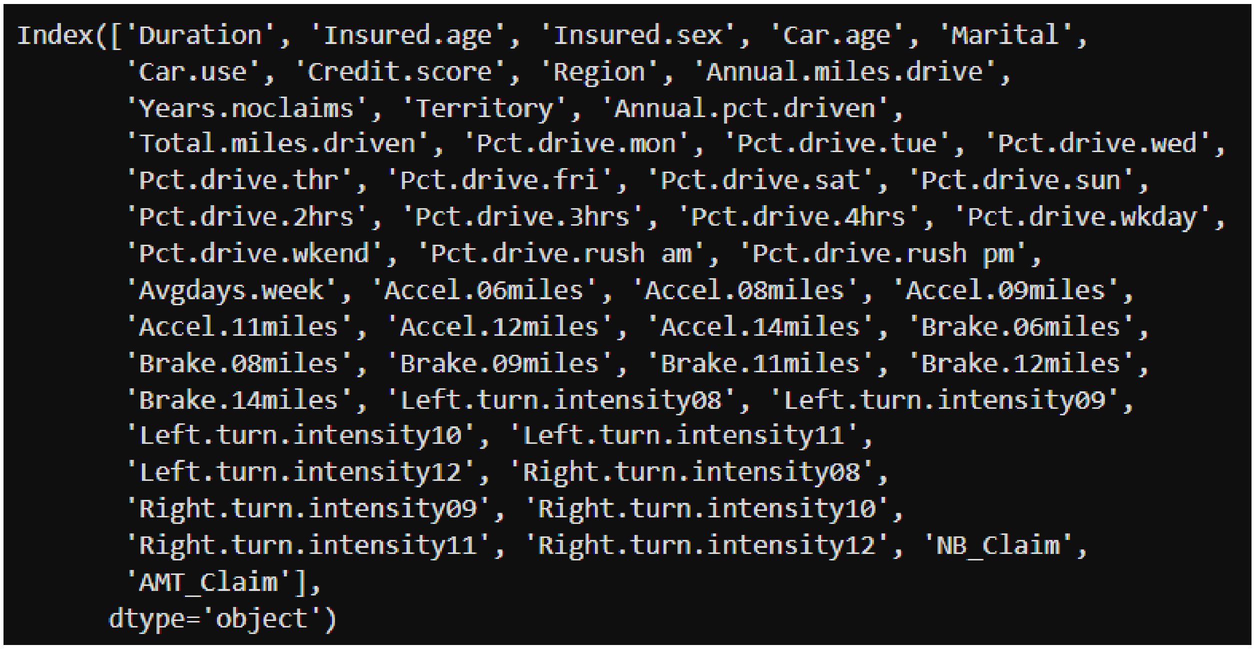
Report

By Nayandeep Deb

- Data Processing

The first task was to extract a useful and relevant dataset for this task to show useful architecture and results. This was a difficult task as there are very few datasets made for this specific topic.

I had finally concluded to use a model with the following features despite the dataset being made largely of SMOTE data which will cause the accuracy of the model to decrease greatly.



```
Index(['Duration', 'Insured.age', 'Insured.sex', 'Car.age', 'Marital',
       'Car.use', 'Credit.score', 'Region', 'Annual.miles.drive',
       'Years.noclaims', 'Territory', 'Annual.pct.driven',
       'Total.miles.driven', 'Pct.drive.mon', 'Pct.drive.tue', 'Pct.drive.wed',
       'Pct.drive.thr', 'Pct.drive.fri', 'Pct.drive.sat', 'Pct.drive.sun',
       'Pct.drive.2hrs', 'Pct.drive.3hrs', 'Pct.drive.4hrs', 'Pct.drive.wkday',
       'Pct.drive.wkend', 'Pct.drive.rush am', 'Pct.drive.rush pm',
       'Avgdays.week', 'Accel.06miles', 'Accel.08miles', 'Accel.09miles',
       'Accel.11miles', 'Accel.12miles', 'Accel.14miles', 'Brake.06miles',
       'Brake.08miles', 'Brake.09miles', 'Brake.11miles', 'Brake.12miles',
       'Brake.14miles', 'Left.turn.intensity08', 'Left.turn.intensity09',
       'Left.turn.intensity10', 'Left.turn.intensity11',
       'Left.turn.intensity12', 'Right.turn.intensity08',
       'Right.turn.intensity09', 'Right.turn.intensity10',
       'Right.turn.intensity11', 'Right.turn.intensity12', 'NB_Claim',
       'AMT_Claim'],
      dtype='object')
```

This model was originally used for a classification task on predicting NB_Claim, but I repurposed it for this task to a regression-based task of predicting Credit score.
https://spectrum.library.concordia.ca/id/eprint/992746/1/Alipanah_MASc_F2023.pdf

This publication was a good read to point me in the direction I should proceed with the model-making part of the project. I also performed Telematic data analysis to find correlations in the data and to extract what features would be useful for this task.

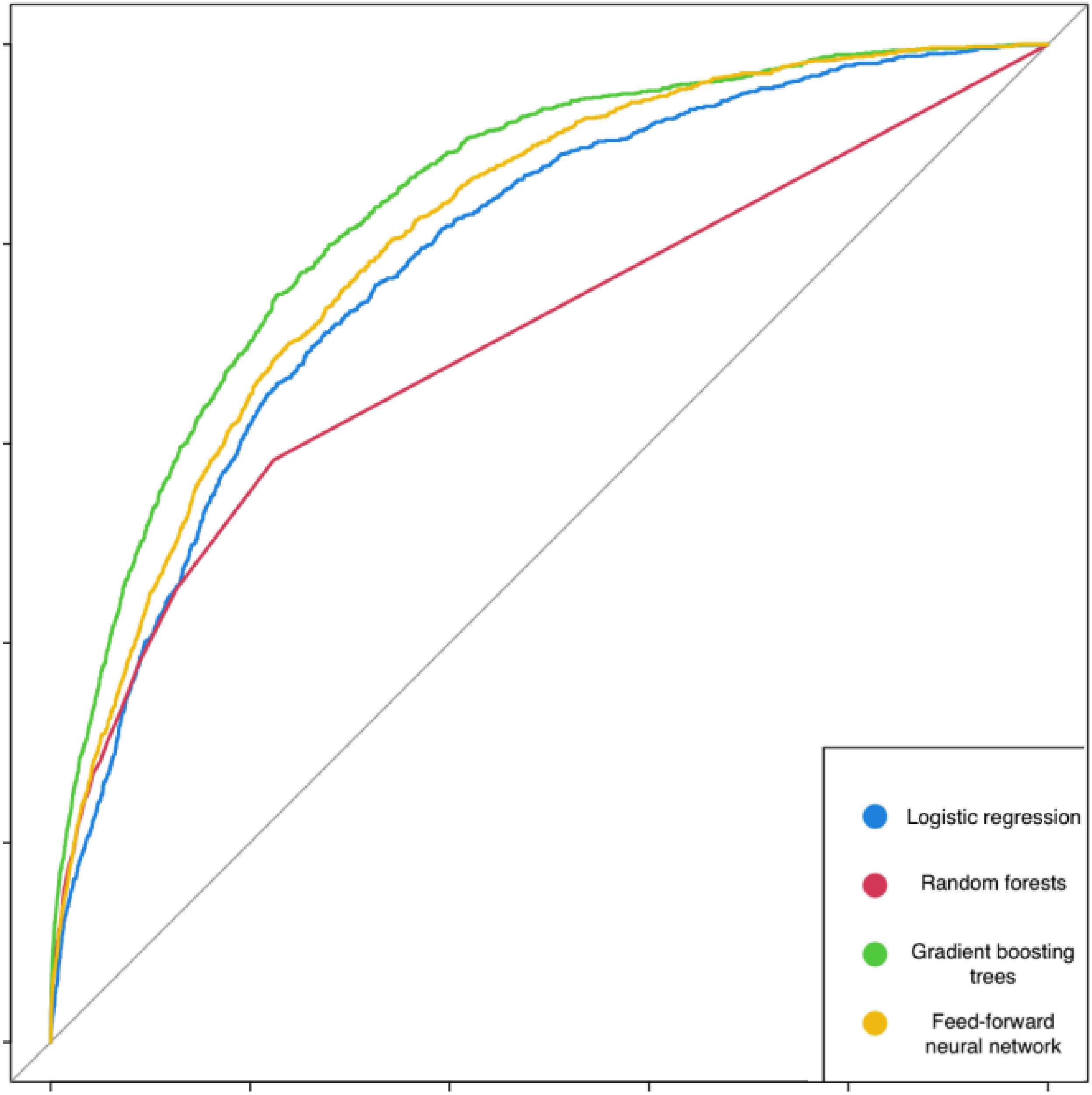
- Data Processing

Other than data processing, I also performed basic data processing (Encoding, New feature generation, dropping of pointless/redundant features, etc.)

This pre-processing has been done directly in the class where Credit score predictions are being made.

- Risk Scoring Model

For the synthetic dataset that was used in this project, Tree-based models give the best output as can be seen in the below graph.



Due to the greatly synthesized data, the accuracy of the data with the labels might not have strong links, so using a Neural Network-based model can actually give poor result with a higher degree of generalization, leading to similar values for all rows.

For a dataset with a high degree of randomness, tree-based models provide greater accuracy for expected values.

For real-life data we can expect greater set of relevant features like geographical features and weather-based features. These more realistic results can bring a greater degree of relatability between features and the credit score. Based on analysis of the current dataset, it is reasonable to assume that either boosting models or Neural Network models will give the best performance, and hands-on data analysis will give us more concrete metrics to decide on which is the best.

- Pricing Engine

The pricing Engine I have currently employed is slightly dynamic but only dependent on one variable: credit score. Having a dataset that can analyze credit score, environmental factors, and personal history, to give a more robust price will greatly improve the architecture, but I have not been able to implement that here. The formula I applied is as follows:

$$(\text{Credit Score})=\exp(-\gamma\times(\text{Credit Score}))$$

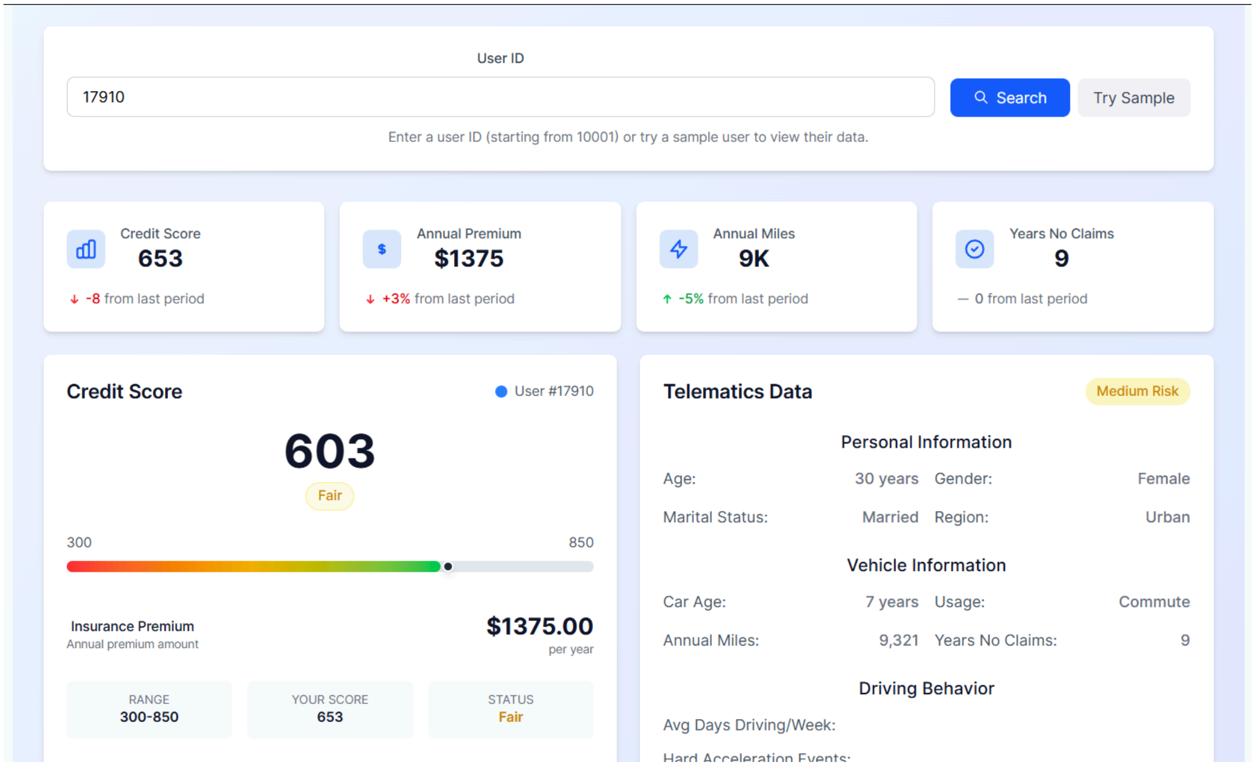
OR

Perform credit score binning to assign it a factor to get multiplied by

- User Dashboard

Prepared a FastAPI-based backend architecture to locally host a server to take user information in real-time (API unavailable), hosting the Tree-Based model that was trained on the dataset, A server having API Endpoints for the frontend to get relevant information from the database when a user wants to access the information.

Prepared a basic dashboard for 10000 users on react. Attached below is an image of the dashboard:



17910

Enter a user ID (starting from 10001) or try a sample user to view their data.

Credit Score 653
↓ -8 from last period

Annual Premium $1375
↓ +3% from last period

Annual Miles 9K
↑ -5% from last period

Years No Claims 9
— 0 from last period

Credit Score ● User #17910

603

Fair

300 850

Insurance Premium
Annual premium amount

$1375.00
per year

RANGE	YOUR SCORE	STATUS
300-850	653	Fair

Telematics Data Medium Risk

Personal Information

Age:	30 years	Gender:	Female
Marital Status:	Married	Region:	Urban

Vehicle Information

Car Age:	7 years	Usage:	Commute
Annual Miles:	9,321	Years No Claims:	9

Driving Behavior

Avg Days Driving/Week:

Hard Acceleration Events:

- Future Improvements and Scalability

1. Integrating weather API features into the TelematicExtraction folder
2. Employing a Mixture of Experts approach to train various models on different subsections of the data to give a more robust prediction on the final dataset.
3. Applying an Authentication System on the Dashboard to only allow valid users to check relevant data
4. Using more tables in the database to store a greater variation of information that can be used for the improvement of dashboard information or style