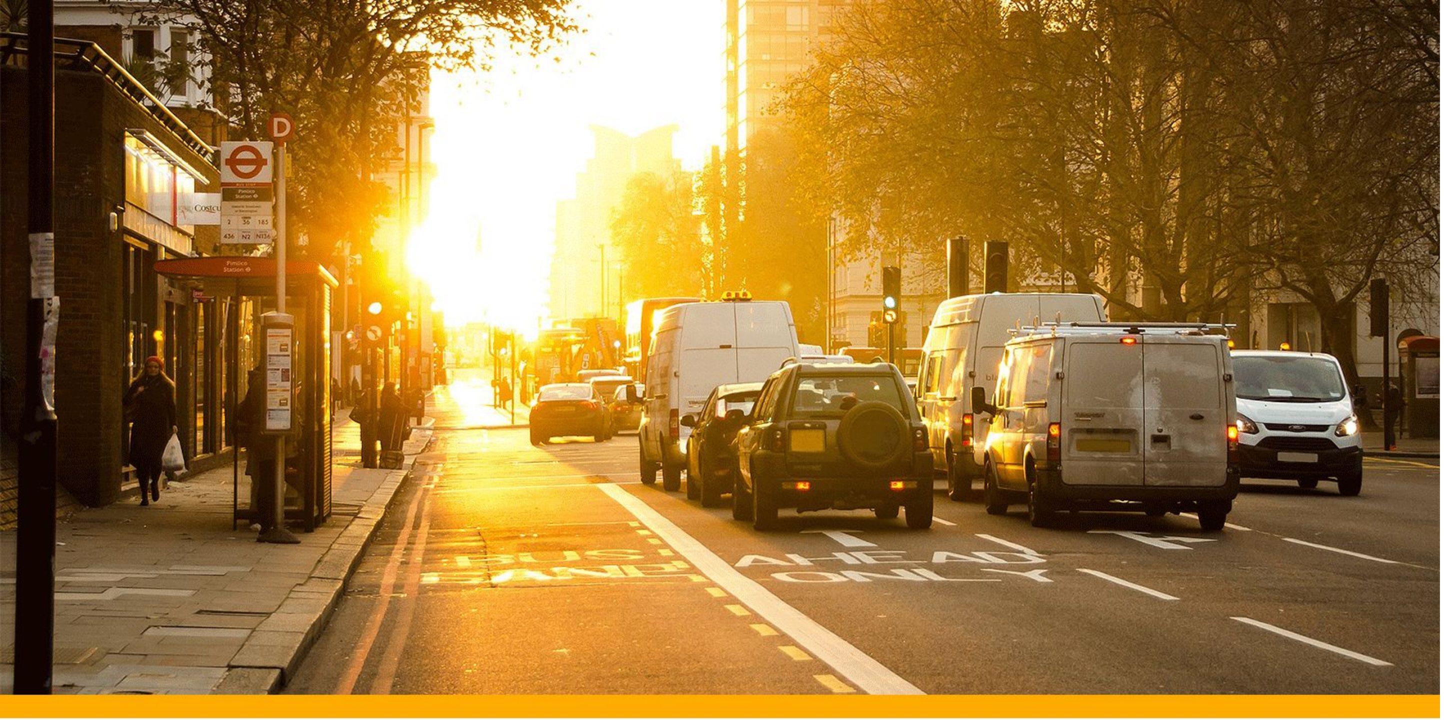


Comparing Traffic Accident Prediction to Codeless AI

CONTENTS

01 Problem Statement

02 About GCP

03 EDA

04 Model

05 Comparison

06 Conclusion

01 Problem Statement

Problem Statement

Big industry players that provided cloud computing services such as Amazon and Google have been releasing codeless AI platform lately and the department is interested in looking whether it is worth expanding their investment in such services provided. Currently, they are in the midst of expanding their infrastructure and looking into acquiring cloud computing services.

With a provided sample data, they would like the following outcome to further expedite their decision:

1. Create a supervised classification model to predict accident severity and compare the outcome together with a codeless AI platform
2. Explore the related services provided and if it's beneficial to the team

02 About Google Console Platform

Google Cloud Platform

- **Google Cloud Platform** (**GCP**), offered by Google, is a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products.
- They provides infrastructure as a service , platform as a service , and serverless computing environments.
- Has total of over 100 products under the Google Cloud brand that can be use for computing, storage, networking,Big Data, Cloud AI and etc

Google Services used in this project

Data Studio

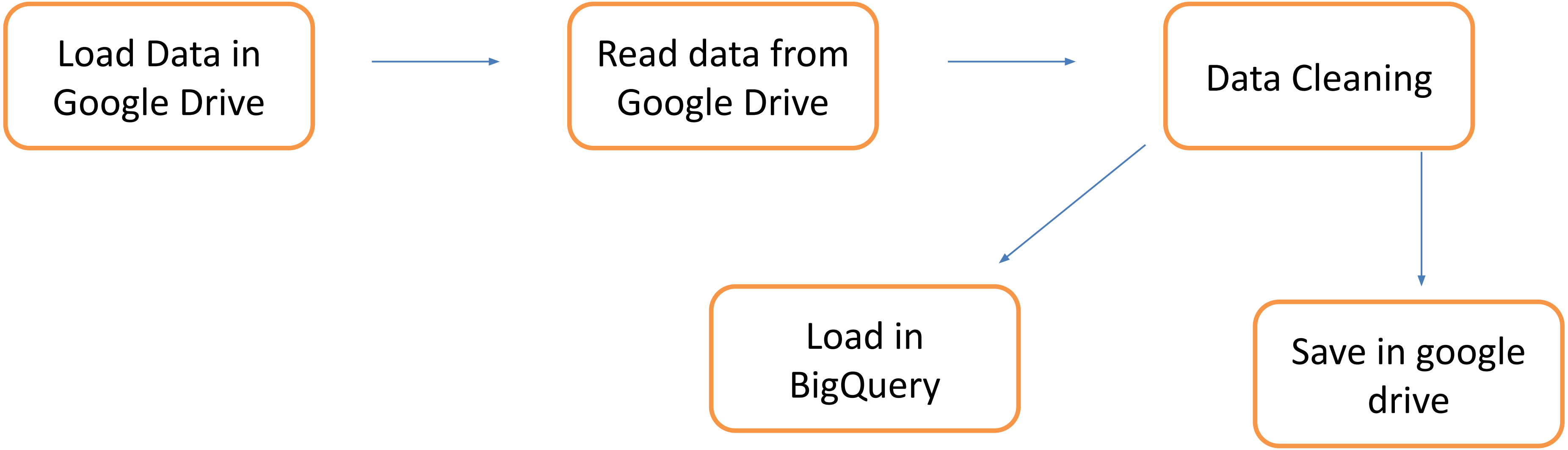
Vertex AI

BigQuery, Google Drive

Colaboratory

03 Data Findings and Summary

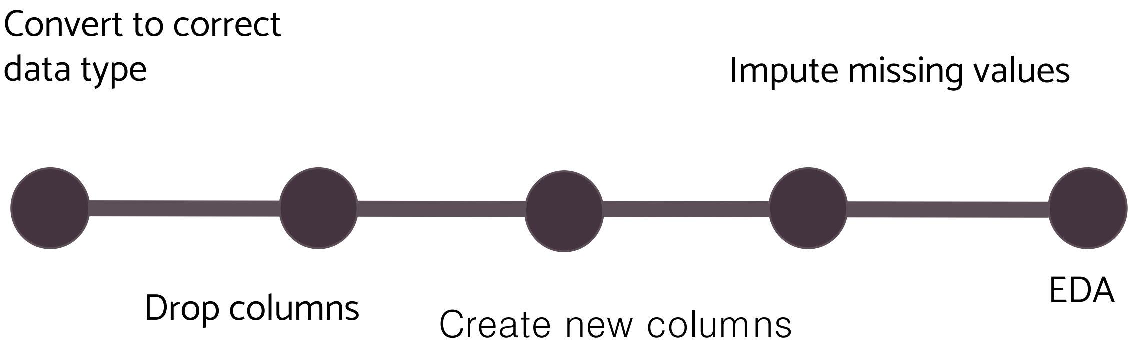
Data Loading Steps:



Data Findings

- Total of 91,199 records and 27 columns
- Missing data are indicated as -1

Data Cleaning



Convert to correct data type

Drop columns

Create new columns

Impute missing values

EDA

Drop columns

- Road_surface_conditions,special_conditions_at_site,carriageway_hazards second_road_class column data mostly belongs to “None”
- junction_control have more than 50% null values

New columns created

1. Season
   - Based on day of year
2. Month
3. Hour
4. Co-ordinates
   - From latitude,longitude

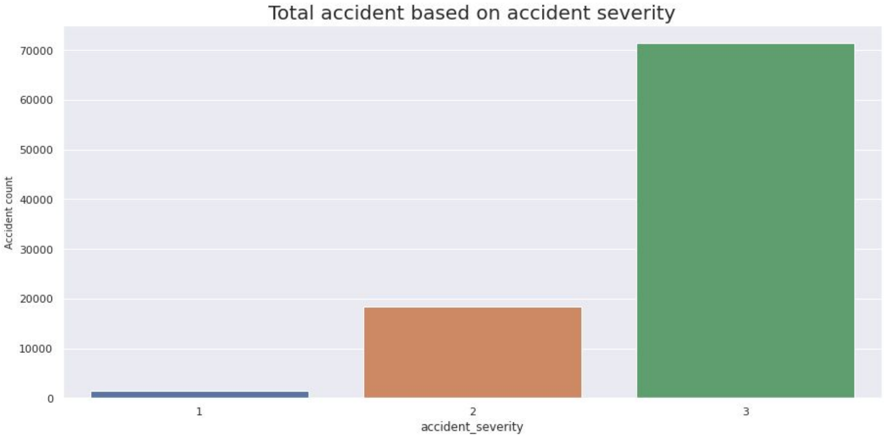
Missing Values

1. Speed limit
   - 10/12 of this null values belongs occurs in the urban area
   - Accidents found in urban area speed found to be mostly in speed limit of 30km/h, Rural areas at 60km/h
   - Impute accordingly based on the urban or rural areas
2. Light conditions
   - Use the date and hour to get light conditions
3. Junction detail
   - Uses median value to impute

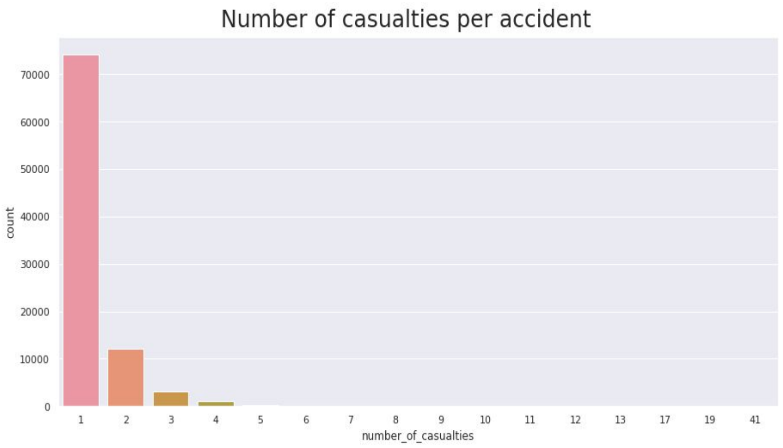
4 . Ignore latitude and longitude as won’t be considered in model but to be used for data visualization

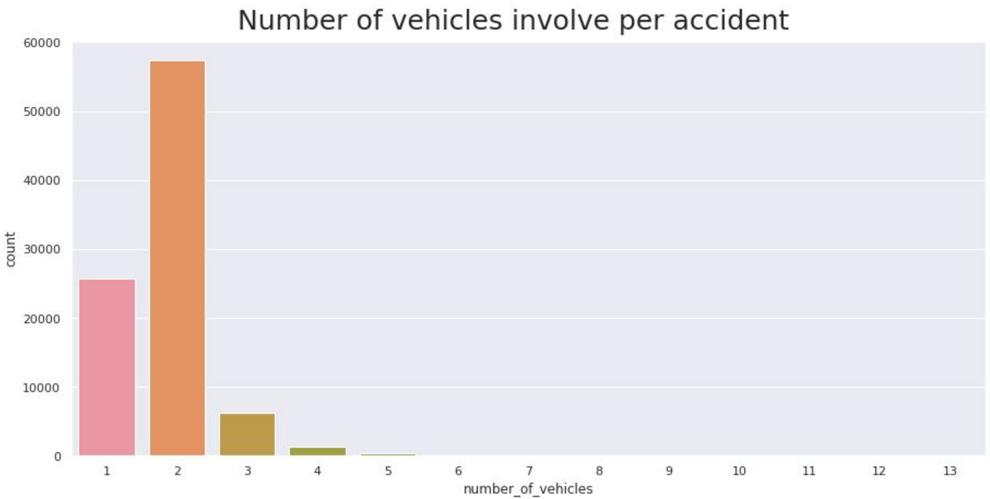
03 EDA

Checking of data imbalance

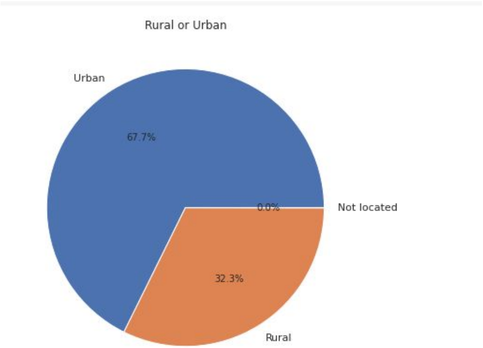


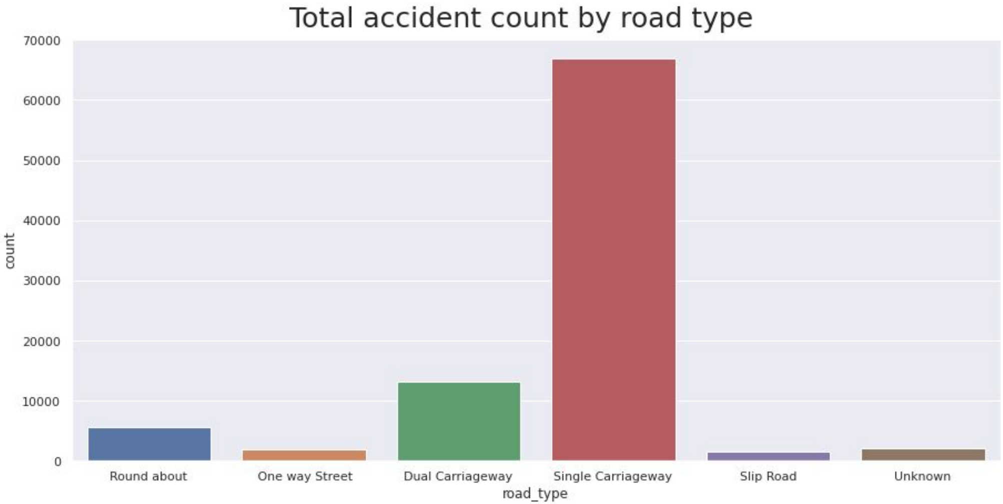
Casualties and vehicles involved



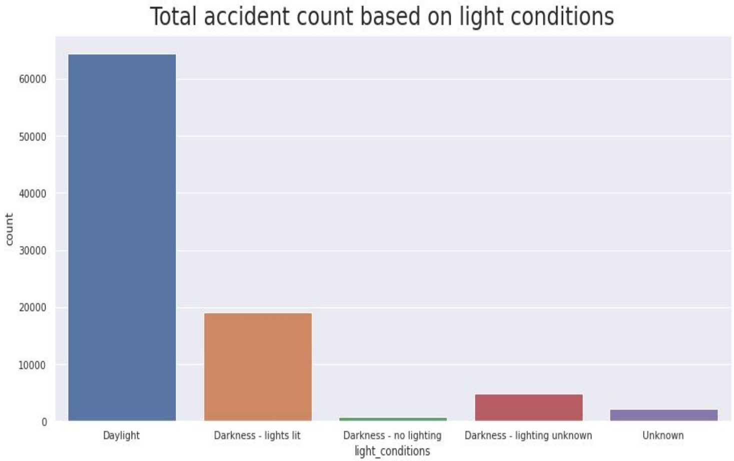


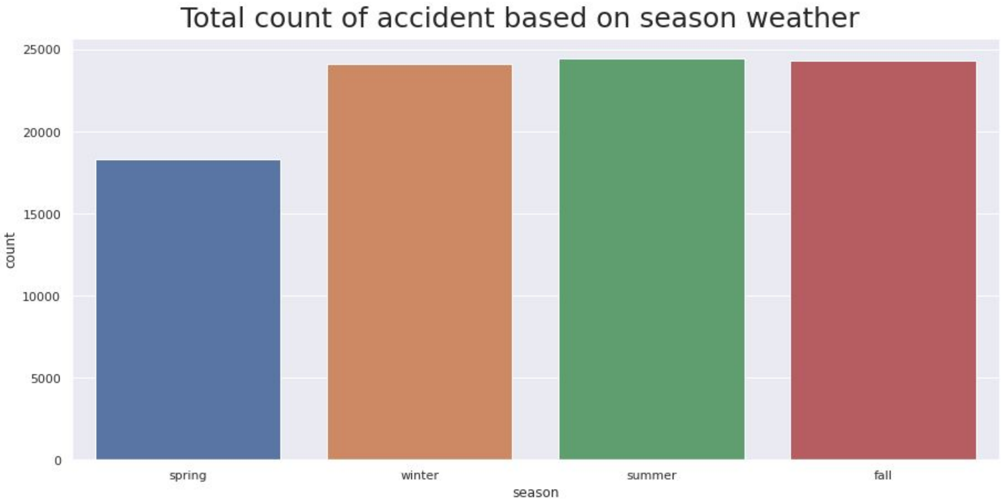
Road type and Urban or Rural



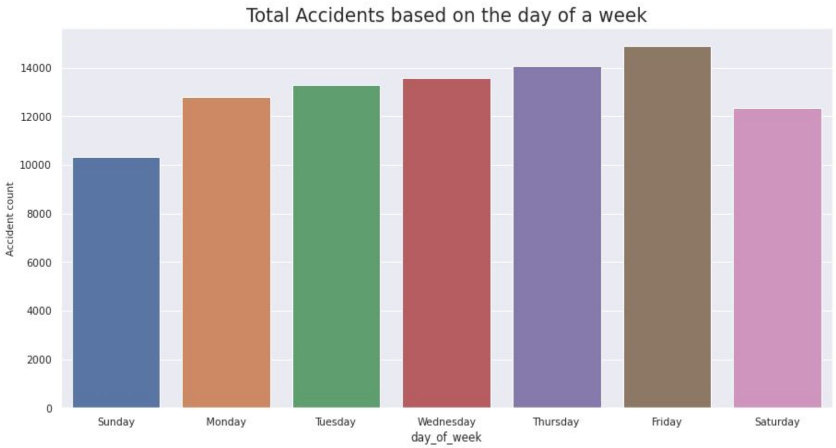


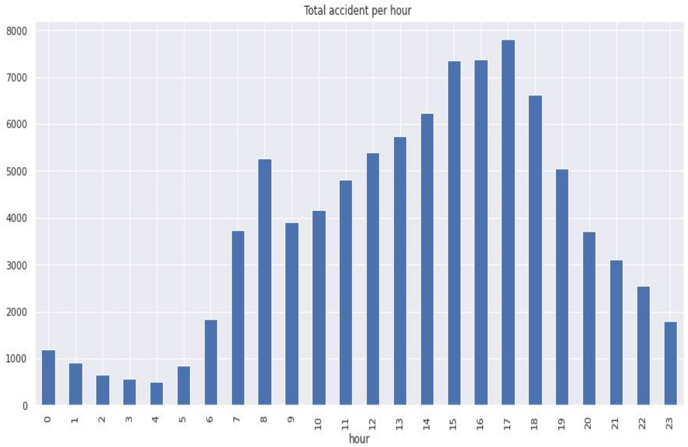
Light conditions and season





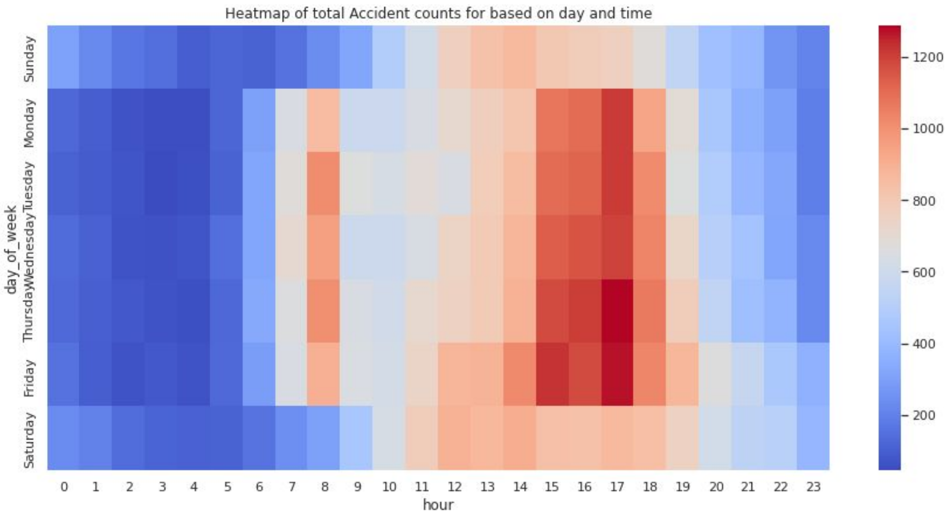
Time and Day

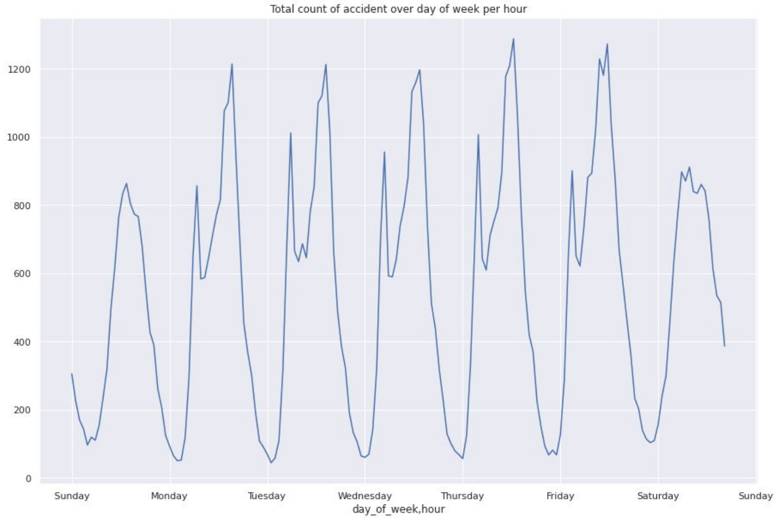




- Most accidents occur generally on Fridays
- Evening peak hour tend to have more accidents than morning peak hour

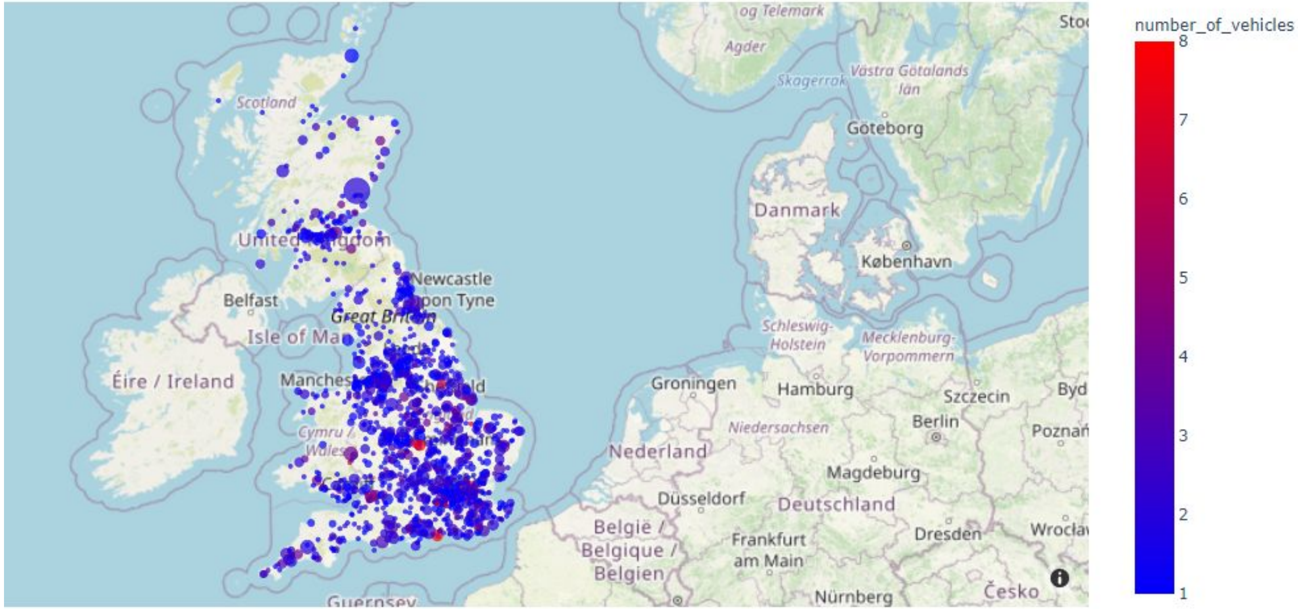
Time and Day





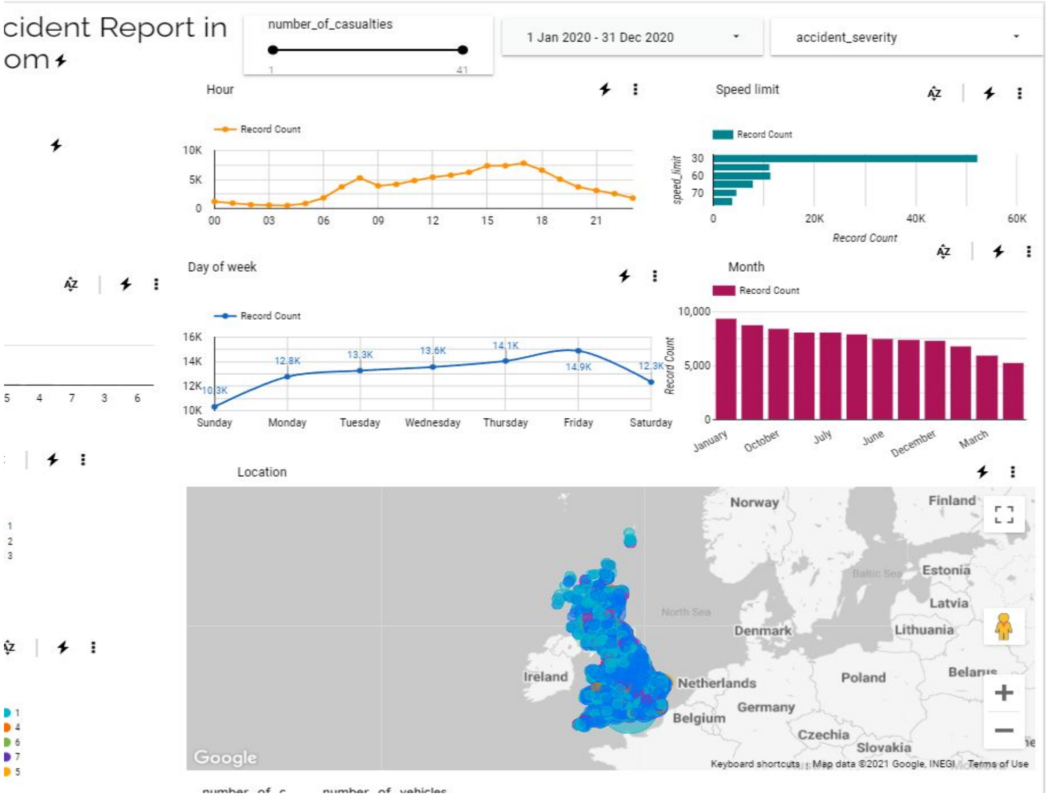
- Most accident occur between 4–6PM especially on
thursdays and fridays

Location of accident severity is fatal



- More accidents occur in london
- Those accidents that involve with more vehicles tend to be on the outskirts or highway

Dashboard

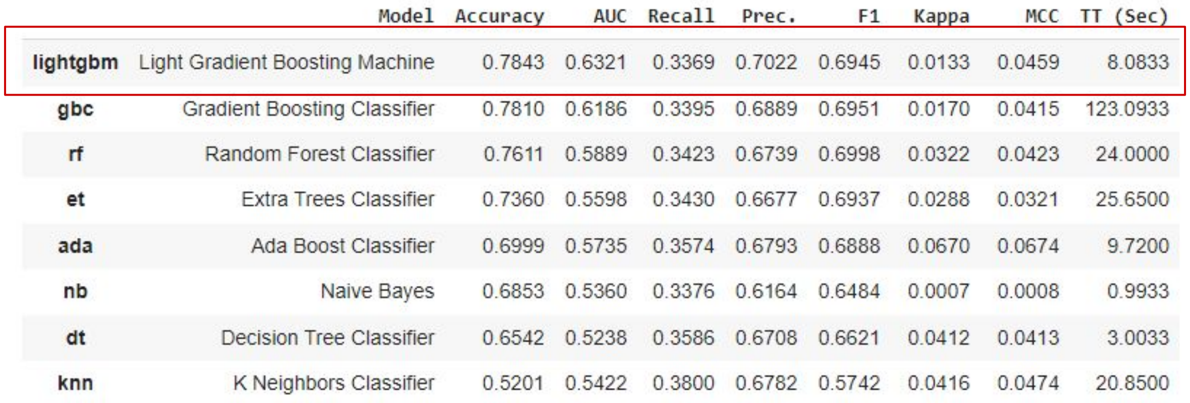


cident Report in
om
number_of_casualties
1
41
1 Jan 2020 - 31 Dec 2020
accident_severity
Hour
Record Count
10K
5K
0
00
03
06
09
12
15
18
21
Speed limit
Record Count
speed_limit
30
60
70
0
20K
40K
60K
Record Count
Day of week
Record Count
16K
14K
12K
10K
10.3K
12.8K
13.3K
13.6K
14.1K
14.9K
12.3K
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Month
Record Count
10,000
5,000
0
Record Count
January
October
July
June
December
March
5
4
7
3
6
1
2
3
1
4
6
7
5
Location
Norway
Finland
Estonia
Latvia
Lithuania
Baltic Sea
North Sea
Denmark
Ireland
Netherlands
Poland
Belarus
Germany
Belgium
Czechia
Slovakia
Google
Keyboard shortcuts
Map data ©2021 Google, INEGI
Terms of Use
number_of_c...
number_of_vehicles

04
Model

Modelling

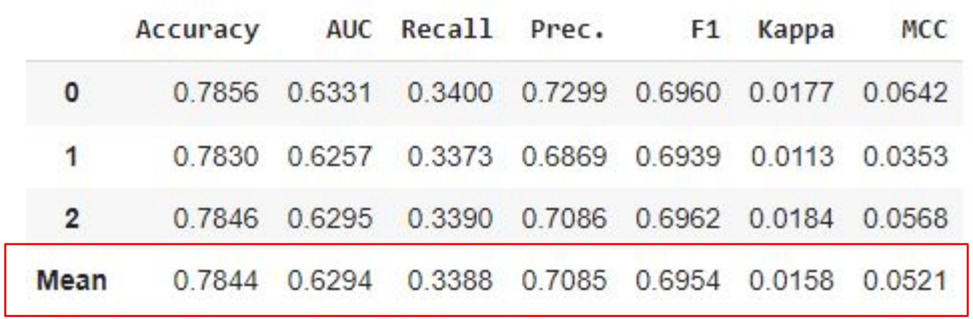
- Multiple factors found which influence the accident severity
- Enable polynomial selection, feature selection, fix_imbalance
- Uses Recall and F1 to select best model



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7843	0.6321	0.3369	0.7022	0.6945	0.0133	0.0459	8.0833
gbc	Gradient Boosting Classifier	0.7810	0.6186	0.3395	0.6889	0.6951	0.0170	0.0415	123.0933
rf	Random Forest Classifier	0.7611	0.5889	0.3423	0.6739	0.6998	0.0322	0.0423	24.0000
et	Extra Trees Classifier	0.7360	0.5598	0.3430	0.6677	0.6937	0.0288	0.0321	25.6500
ada	Ada Boost Classifier	0.6999	0.5735	0.3574	0.6793	0.6888	0.0670	0.0674	9.7200
nb	Naive Bayes	0.6853	0.5360	0.3376	0.6164	0.6484	0.0007	0.0008	0.9933
dt	Decision Tree Classifier	0.6542	0.5238	0.3586	0.6708	0.6621	0.0412	0.0413	3.0033
knn	K Neighbors Classifier	0.5201	0.5422	0.3800	0.6782	0.5742	0.0416	0.0474	20.8500

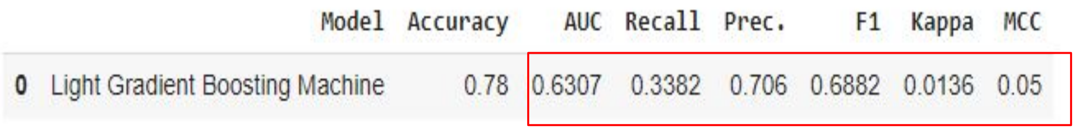
Create and Tune Selected model

- From selected model, input the parameters to tune the model



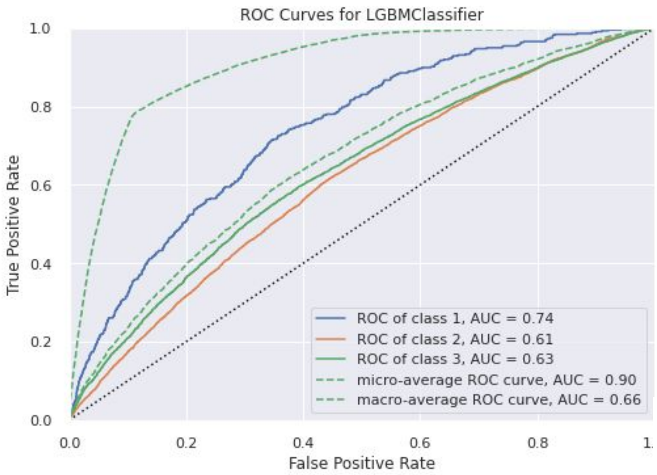
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7856	0.6331	0.3400	0.7299	0.6960	0.0177	0.0642
1	0.7830	0.6257	0.3373	0.6869	0.6939	0.0113	0.0353
2	0.7846	0.6295	0.3390	0.7086	0.6962	0.0184	0.0568
Mean	0.7844	0.6294	0.3388	0.7085	0.6954	0.0158	0.0521

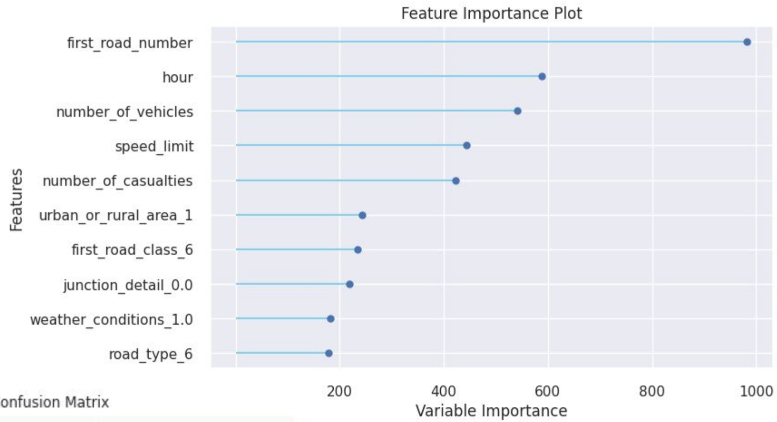
Test Data



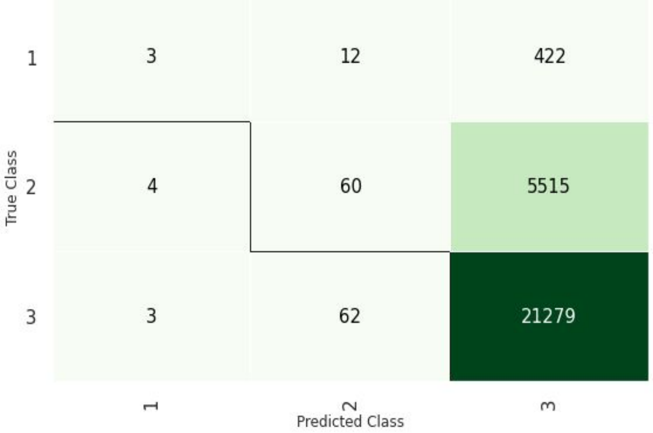
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Light Gradient Boosting Machine	0.78	0.6307	0.3382	0.706	0.6882	0.0136	0.05

Metrics Output


ROC Curves for LGBMClassifier


Feature Importance Plot

LGBMClassifier Confusion Matrix

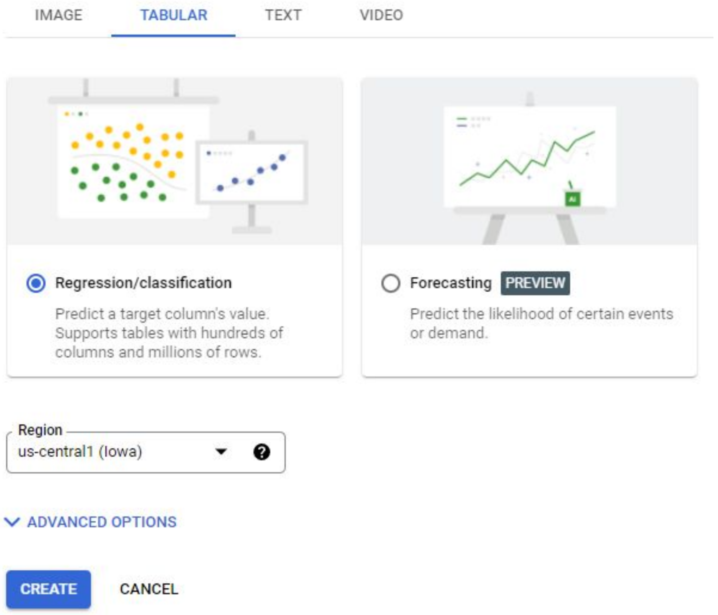


	1	2	3
1	3	12	422
2	4	60	5515
3	3	62	21279

Vertex AI

Datasets

- From the saved csv file in google drive, create a dataset by importing the csv file from google drive and choose the model type



● Regression/classification

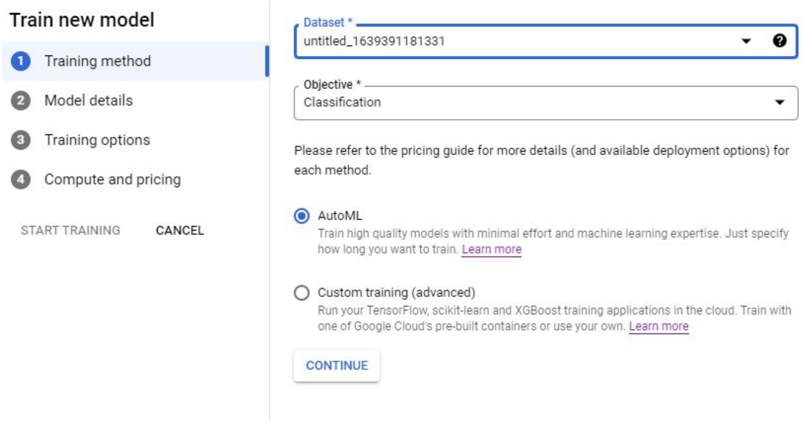
Predict a target column's value. Supports tables with hundreds of columns and millions of rows.

○ Forecasting PREVIEW

Predict the likelihood of certain events or demand.

Region: us-central1 (Iowa)

Vertex AI



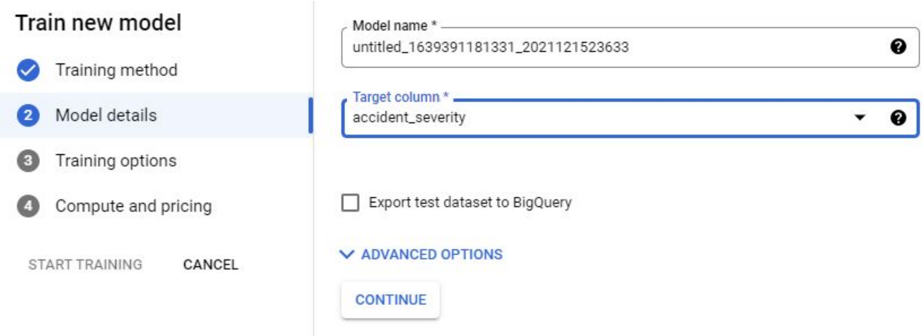
Dataset *: untitled_1639391181331

Objective *: Classification

Please refer to the pricing guide for more details (and available deployment options) for each method.

● AutoML
Train high quality models with minimal effort and machine learning expertise. Just specify how long you want to train. Learn more

○ Custom training (advanced)
Run your TensorFlow, scikit-learn and XGBoost training applications in the cloud. Train with one of Google Cloud's pre-built containers or use your own. Learn more



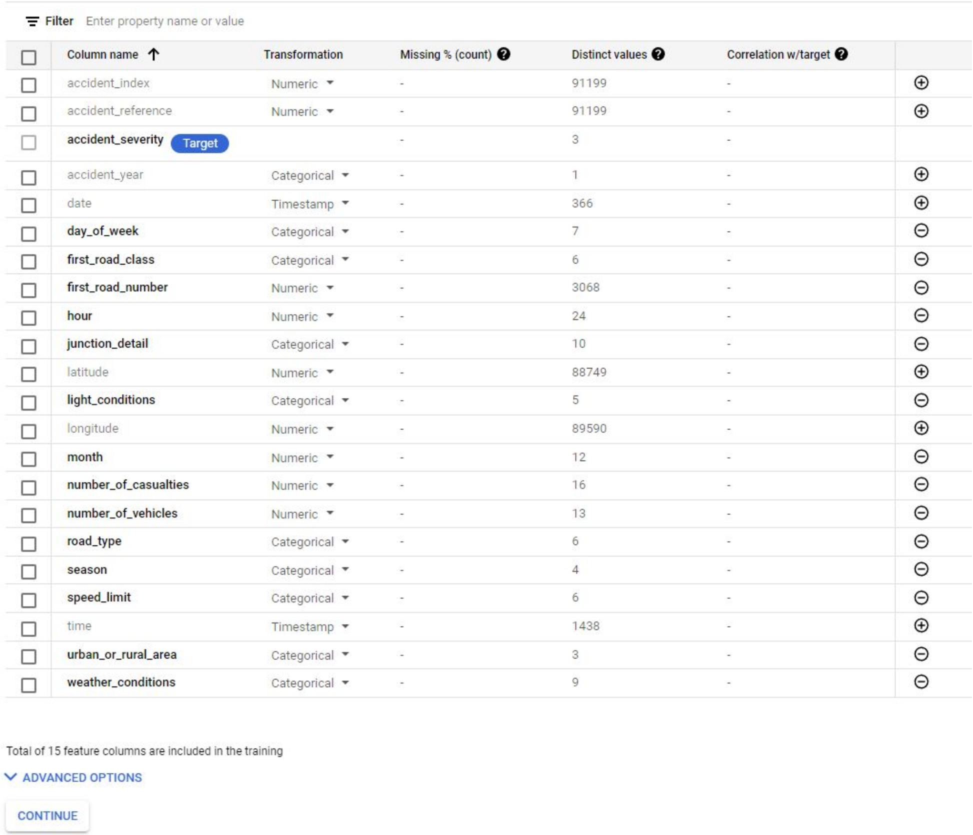
Model name *: untitled_1639391181331_2021121523633

Target column *: accident_severity

□ Export test dataset to BigQuery

- Select AutoML to use Google's codeless AI platform
- Custom training can be selected only for existing deployed python applications in containers

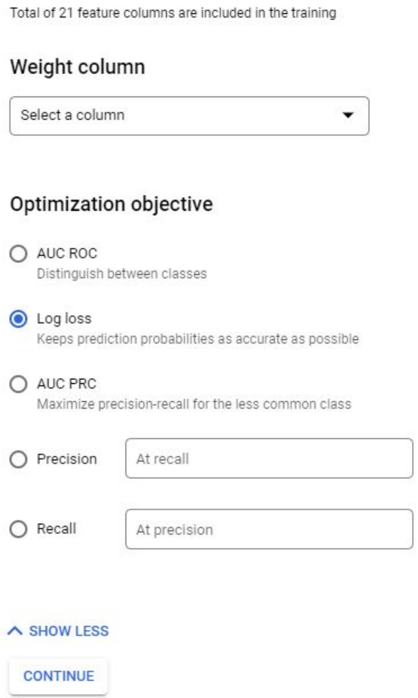
Vertex AI



Filter Enter property name or value

□	Column name ↑	Transformation	Missing % (count)	Distinct values	Correlation w/target	
□	accident_index	Numeric	-	91199	-	⊕
□	accident_reference	Numeric	-	91199	-	⊕
□	accident_severity Target		-	3	-	
□	accident_year	Categorical	-	1	-	⊕
□	date	Timestamp	-	366	-	⊕
□	day_of_week	Categorical	-	7	-	⊖
□	first_road_class	Categorical	-	6	-	⊖
□	first_road_number	Numeric	-	3068	-	⊖
□	hour	Numeric	-	24	-	⊖
□	junction_detail	Categorical	-	10	-	⊖
□	latitude	Numeric	-	88749	-	⊕
□	light_conditions	Categorical	-	5	-	⊖
□	longitude	Numeric	-	89590	-	⊕
□	month	Numeric	-	12	-	⊖
□	number_of_casualties	Numeric	-	16	-	⊖
□	number_of_vehicles	Numeric	-	13	-	⊖
□	road_type	Categorical	-	6	-	⊖
□	season	Categorical	-	4	-	⊖
□	speed_limit	Categorical	-	6	-	⊖
□	time	Timestamp	-	1438	-	⊕
□	urban_or_rural_area	Categorical	-	3	-	⊖
□	weather_conditions	Categorical	-	9	-	⊖

Total of 15 feature columns are included in the training



Total of 21 feature columns are included in the training

Weight column

Select a column

Optimization objective

○ AUC ROC
Distinguish between classes

● Log loss
Keeps prediction probabilities as accurate as possible

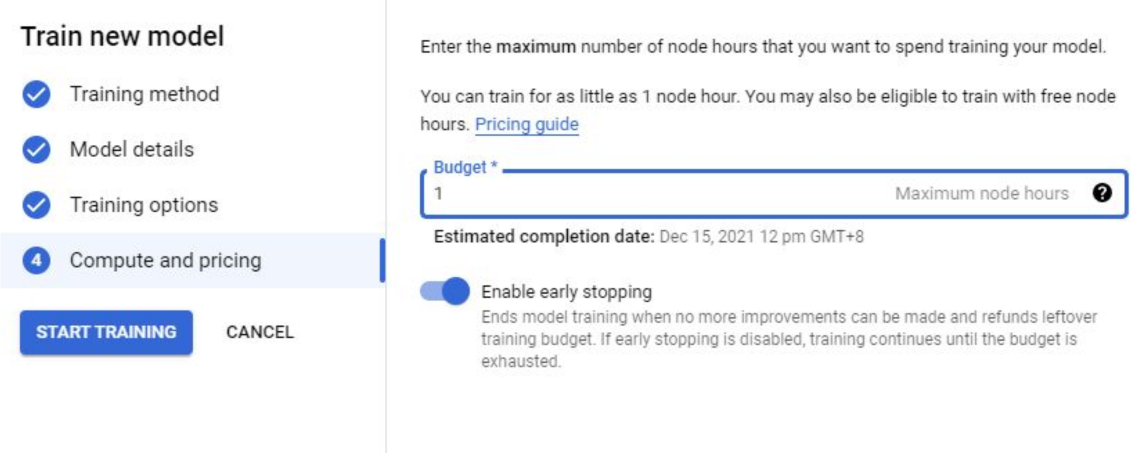
○ AUC PRC
Maximize precision-recall for the less common class

○ Precision At recall

○ Recall At precision

Select features for the model

Vertex AI



Enter the **maximum** number of node hours that you want to spend training your model.

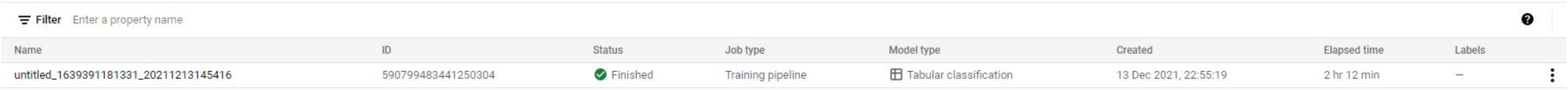
You can train for as little as 1 node hour. You may also be eligible to train with free node hours. Pricing guide

Budget *: 1 Maximum node hours

Estimated completion date: Dec 15, 2021 12 pm GMT+8

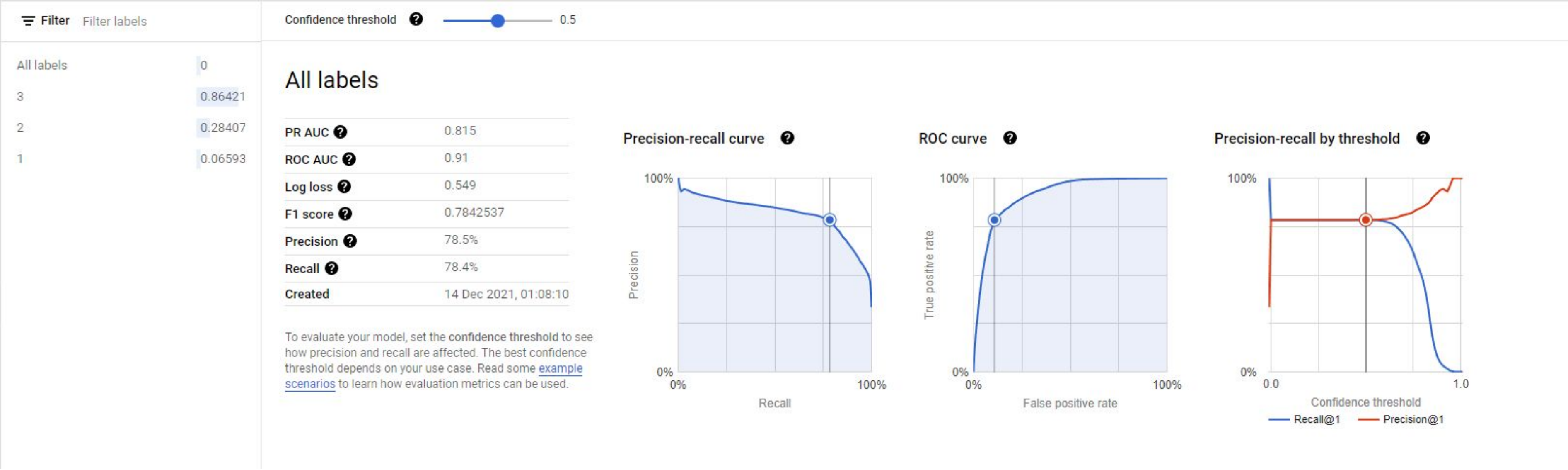
● Enable early stopping

Ends model training when no more improvements can be made and refunds leftover training budget. If early stopping is disabled, training continues until the budget is exhausted.



Name	ID	Status	Job type	Model type	Created	Elapsed time	Labels
untitled_1639391181331_20211213145416	590799483441250304	✓ Finished	Training pipeline	Tabular classification	13 Dec 2021, 22:55:19	2 hr 12 min	–

Vertex AI



Filter Filter labels

All labels	0
3	0.86421
2	0.28407
1	0.06593

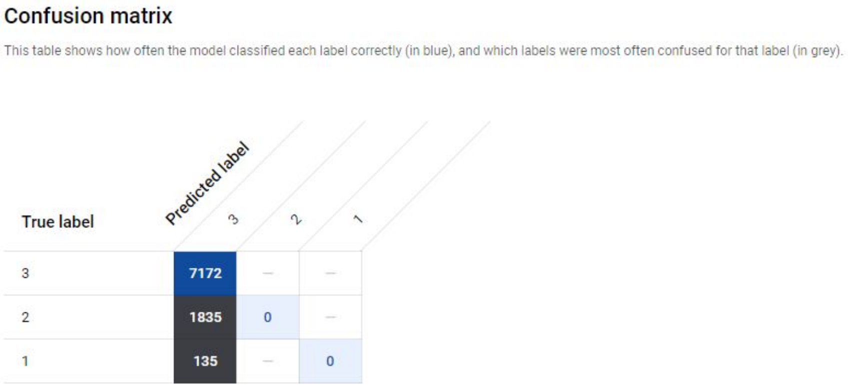
Confidence threshold 0.5

All labels

PR AUC	0.815
ROC AUC	0.91
Log loss	0.549
F1 score	0.7842537
Precision	78.5%
Recall	78.4%
Created	14 Dec 2021, 01:08:10

To evaluate your model, set the **confidence threshold** to see how precision and recall are affected. The best confidence threshold depends on your use case. Read some example scenarios to learn how evaluation metrics can be used.

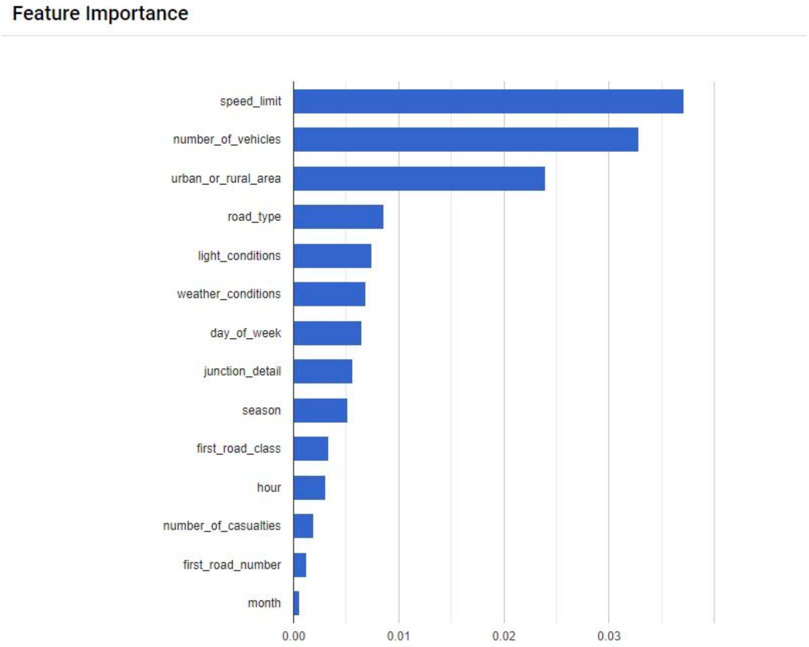
Vertex AI



Confusion matrix

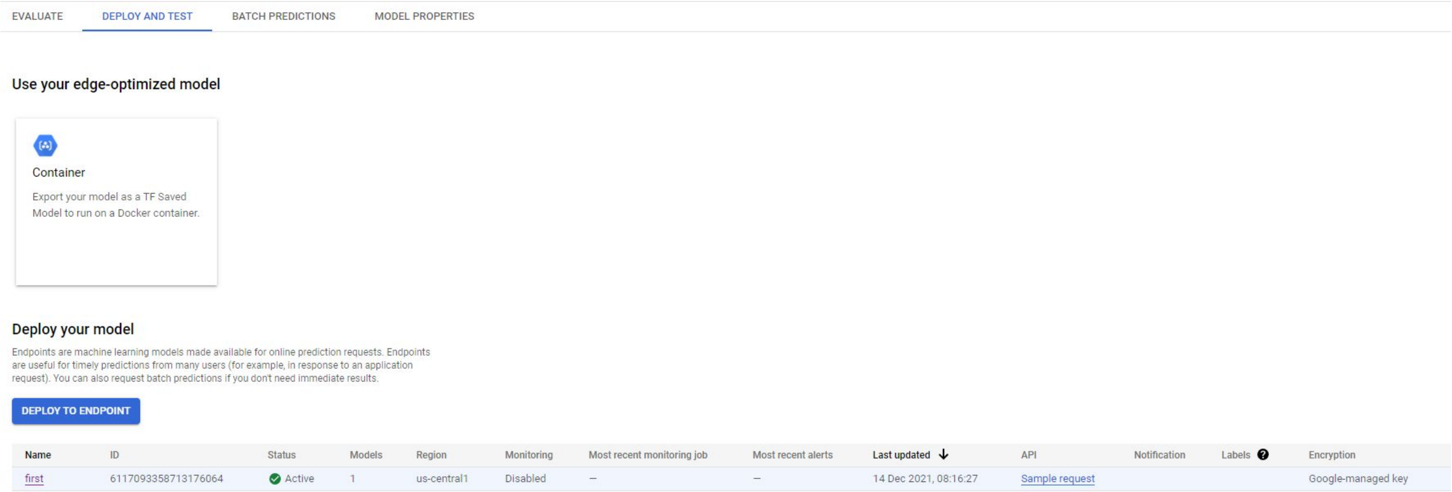
This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in grey).

True label	Predicted label 3	2	1
3	7172	—	—
2	1835	0	—
1	135	—	0



Feature Importance

Deploy and test model



Use your edge-optimized model

Container

Export your model as a TF Saved Model to run on a Docker container.

Deploy your model

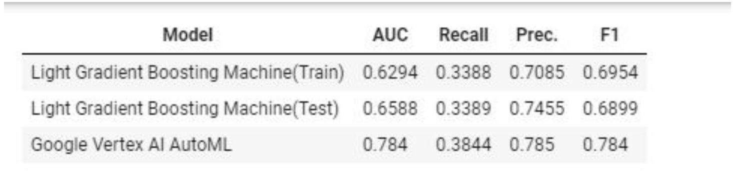
Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

Name	ID	Status	Models	Region	Monitoring	Most recent monitoring job	Most recent alerts	Last updated ↓	API	Notification	Labels ?	Encryption
first	6117093358713176064	✓ Active	1	us-central1	Disabled	–	–	14 Dec 2021, 08:16:27	Sample request			Google-managed key

- Can choose to deploy and test predictions online for many users

05 Comparison

Comparison of results



Model	AUC	Recall	Prec.	F1
Light Gradient Boosting Machine(Train)	0.6294	0.3388	0.7085	0.6954
Light Gradient Boosting Machine(Test)	0.6588	0.3389	0.7455	0.6899
Google Vertex AI AutoML	0.784	0.3844	0.785	0.784

- Loss between the train and test chosen model is very little for the created model
- Google Vertex AI AutoML has higher precision and f1 score comparing to the chosen model.
- However, there are disadvantages in using such codeless model

Findings of using Vertex AI AutoML

1. Lack of customization and control on what goes into the model
2. No transparency what was done by the model
3. Unable to use hyperparameter tuning to further enhance
4. Unable to enable properties such as feature selection,polynomial features or SMOTE
5. Limits to data size at 1M
6. For classification prediction, only minimize-log-loss is support to optimize which may not always support a project's objective
7. For imbalance class, best practice stated was to have at least 100 rows of data for every class and assign a manual split to make sure enough rows with the minority outcomes are included in every split.

05 Conclusion

Conclusion and Recommendations

- Cloud computing services such as Google cloud platform provide many services that allow users retrieve data online and collaborate to perform business activities
- However, there are limitation using such codeless services for machine learning despite the ease of creating models by just few clicks

Suggestion

- Continue using python libraries available to create predicting models that allows transparency, control and enhancement to the model that may better fit the objective of project
- Deploy them in a cloud computing service so that similar experiment does not have to be re-created and users can use the models at one time to predict or collaborate
- Cloud computing services does benefit the analytics team and it is worth the investment as many related services can be found in one platform thus beneficial to the data analytics team

THANK YOU!