
Mini Project: ETA Prediction

Alireza Kazemi

Overview

ETA Prediction

Our goal is to minimize the error of ETA

Metric

$$MAPE = \frac{|ETA - ATA|}{ATA}$$

Current
About

32.5%

Status
error

Tools

A second model (Regression) corrects the
initial eta

Result

About 22.5% error

Steps to go

Data Insight

- Data Schema
- Features' Insight
- Preprocessing
- Feature Selection
- Feature Engineering

Data Modeling

- Model Selection
 - Optimization & Regularization
 - Model Tuning
 - Model Evaluation
 - Model Explanation
-

Data Insight

Schema

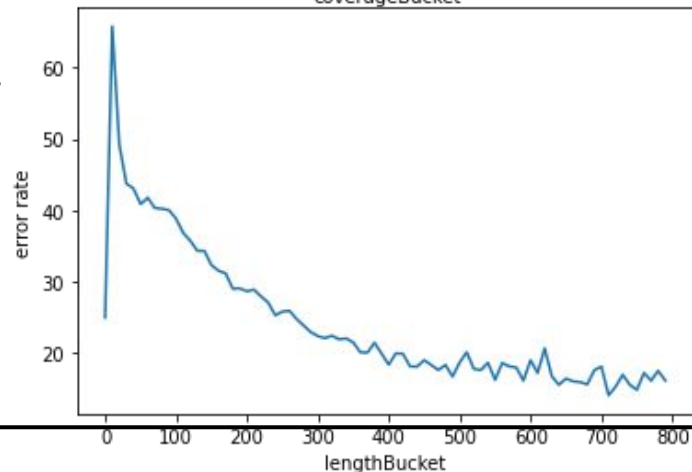
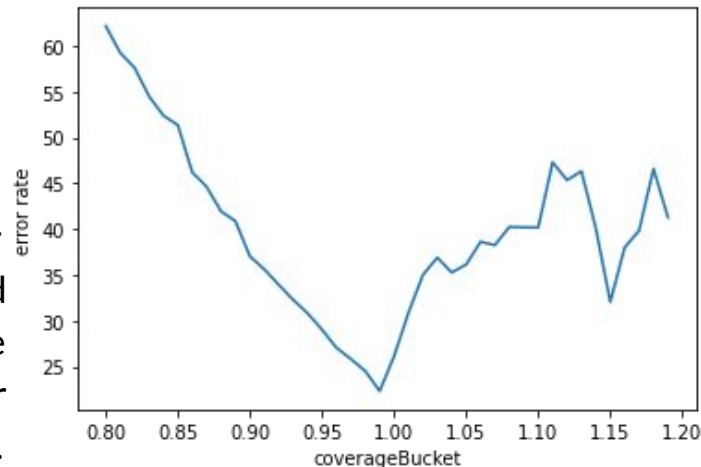
- Length
- ETA
- Coverage
- Lat
- Long
- Timestamp
- Way Segments Count
- Way Type
- isTunnel
- ATA (Target)

Overview

- Train: 160'000 rows
 - Test: 70'000 rows
 - Other Features:
 - Edge ID
 - Name
 - User Group
-

Data Insight

- **Coverage:** Coverage of about 10% of data is lower than 0.85 and higher than 1.05. Not having a good enough coverage causes higher errors in the estimation of each road. Also, we can see a linear function for error rate and coverage (coverage < 1).
- **Length:** Average is 0.2 Km and we can see a linear function for error rate and length.



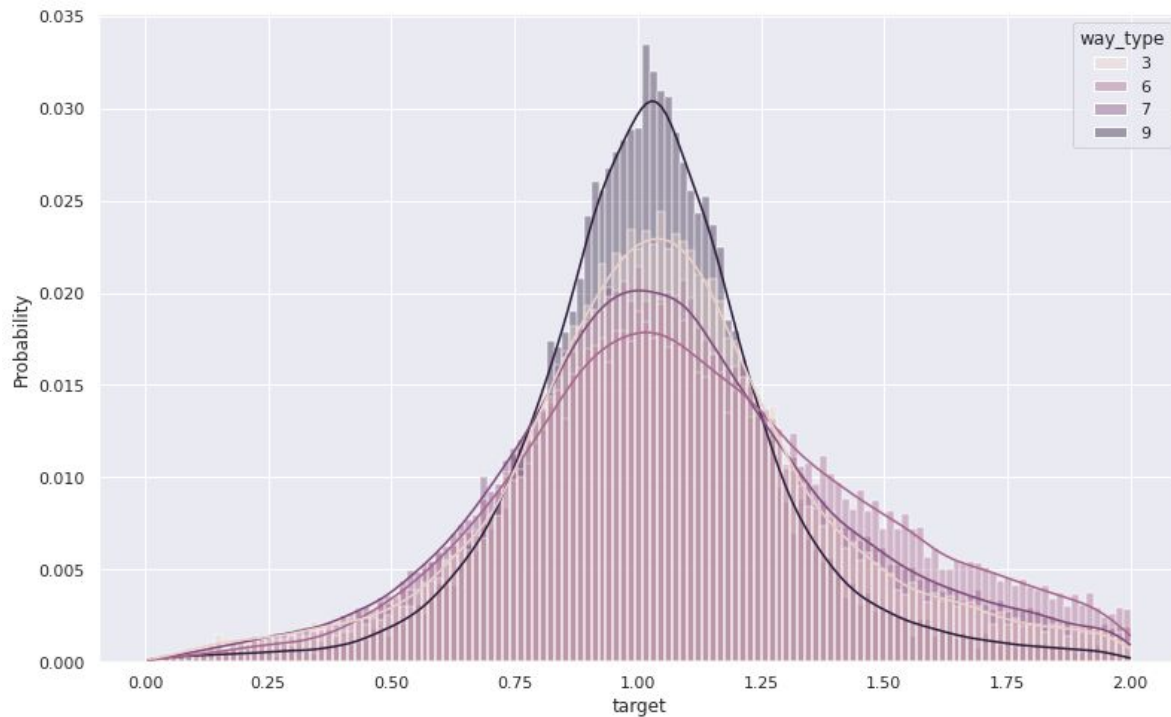
Data Insight

- **Speed:** I calculated each road's speed by dividing length on ATA and it is better to remove high speeds. Only about 1% of data has a speed more than 115 Km/hour.
- **isTunnel:** 0.02% of roads are in tunnels
- **Way Segments Count:** Error rate didn't differ by any different segment count.

```
.withColumn('speed', F.col('length')  
/F.col('ata')*3.6)  
  
.filter(F.col('speed')<=150)  
  
.filter(F.col('coverage').between(0  
.85,1.05))  
  
.filter(F.col('is_tunnel')==0)  
  
.filter(F.col('length')<=2500)
```

Data Insight

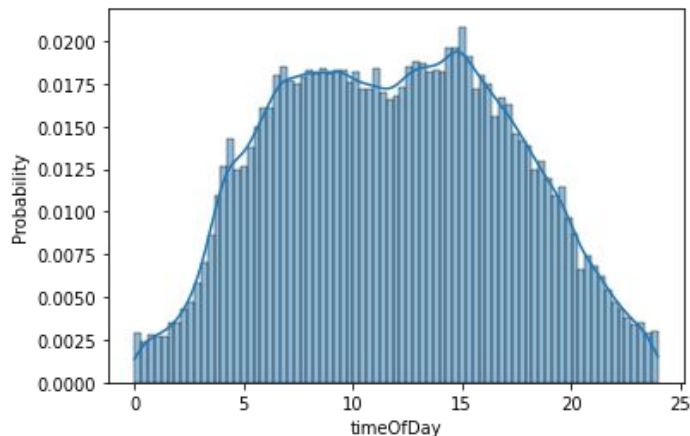
- **WayType:** Way-type of about 98.2% of data is in [3,6,7,9]. We can see way type '6' creates overestimation but others have a random impact.



Way Type	Share	Error
7	42.3%	38%
3	28.3%	34%
6	14.9%	43%
9	12.7%	22%
Others	1.8%	-

Data Insight

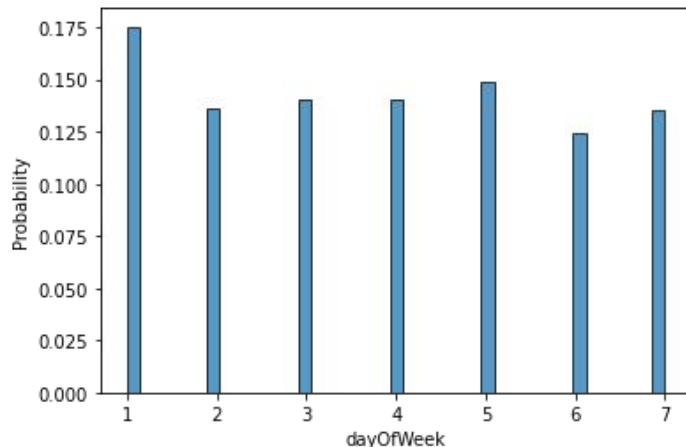
- **Timestamp:** I extracted two features from column 'time' to have a better knowledge more than unix timestamp:
 - Time of Day : Hour + (Minute/60)
 - Day of Week



```
.withColumn('createdAt',F.from_unix  
time(F.col('timestamp')/1000))
```

```
.withColumn('timeOfDay',F.hour('cre  
atedAt')+F.round(F.minute('createdA  
t')/60,2))
```

```
.withColumn('dayOfWeek',F.dayofweek  
('createdAt'))
```



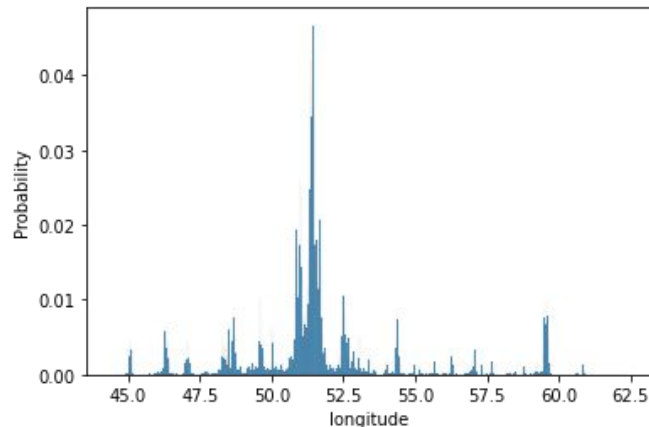
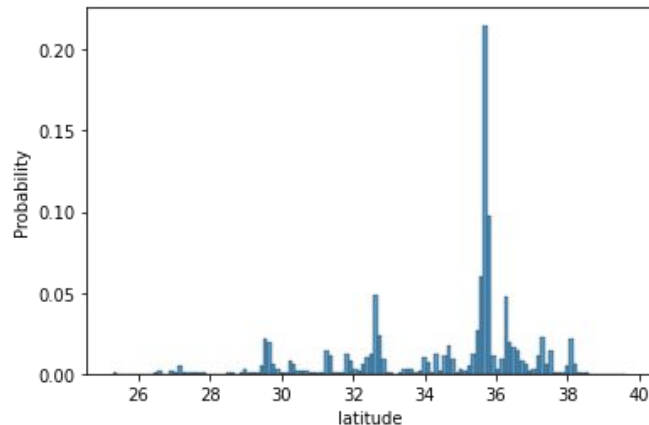
Data Insight

- **Latitude and Longitude:** As we can see, the roads are placed on different locations of Iran. So we can add a feature (zone or isTehran or isBigCity).

```
.withColumn('isInTehran', ((F.col('latitude').between(35.58,35.85)) &  
    (F.col('longitude').between(51.1,51.6))).cast('integer'))
```

- 25% of data are in Tehran. Error rate in Tehran is more than other locations in Iran. So the location may affect the estimation, but it's not necessary to label it, because we use trees for regression.

+-----+-----+	
isInTehran error rate	
+-----+-----+	
0	33.48
1	41.01
+-----+-----+	



Data Modeling

Questions

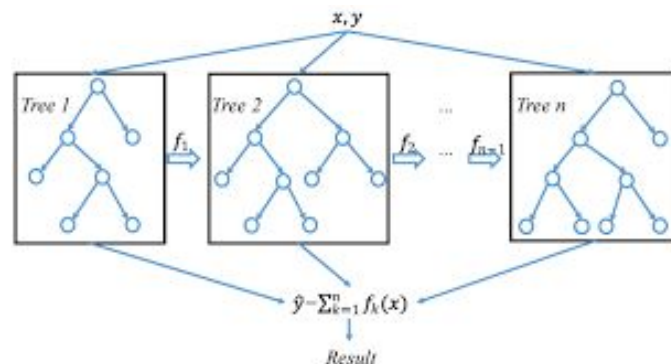
- What are the final features?
 - How to evaluate the model ? (loss function, validation data)
 - What is the best metric for Target?
 - $|ETA - ATA| / ATA$
 - $|ATA - ETA|$
 - ATA
 - What's the best regression model for estimating?
 - Test and train have a same feature distribution?
-

Data Modeling

Model Selection

- NonLinearity dependencies in features (such as time, location)
- The features are not completely independent.
- Our features are few, and, lasso and ridge will causes the weight shrinkage.
- We have enough data for training by random forests or same models.
- The table above, Shows that we have achieved a better loss compared to others. (SVM had a poor performance)
- I used the [XGBoost](#) Model for training. (other candidates: **Light GBM**, **CatBoost**)

XGBoost	22.5	ETA/ATA
Ridge	29	ATA
Lasso	33	ATA
SVM	-	-



Data Modeling

```
.withColumn('target', F.col('eta')/F.col('ata'))
```

```
.withColumn('targetError', F.abs(100*((F.col('eta')-F.col('ata'))/F.col('ata'))))
```

Evaluation

- I splitted data into two parts (10% and 90%) for training and validation in model training process. The model is evaluated by the **MAE of validation data**. I also used early stopping (**25 Rounds**) to prevent overfitting. I evaluated by MAE due to my experience as a customer, the relative error of a forecasting matters to me.

```
bestModel.fit(train_X, (train_y['target']).values.ravel(), early_stopping_rounds=25,  
eval_metric=["mae"], eval_set=eval_set, verbose=True)
```

- The reported metric is:

$$MAPE = \frac{100 * |ETA - ATA|}{ATA}$$

Data Modeling

Tuning

- Parameters
 - Learning Rate: **0.2**
 - Estimator Count: **270**
 - Maximum Depth: **8**
 - Minimum Child Weight: **0**
- Hyper Parametrization Space
 - Random space with 40 trials

```
space = {'max_depth': hp.quniform("max_depth", 3, 14, 1),  
        'min_child_weight': hp.quniform('min_child_weight', 0, 10, 1),  
        'n_estimators': hp.quniform("n_estimators", 200, 350, 30)  
}
```

```
Stopping. Best iteration:  
[77]  validation 0-mae:0.250344  
validation_1-mae:0.261973
```

```
XGBRegressor(eta=0.2,  
eval_metric='mae', max_depth=8,  
min_child_weight=0,  
n_estimators=270)
```

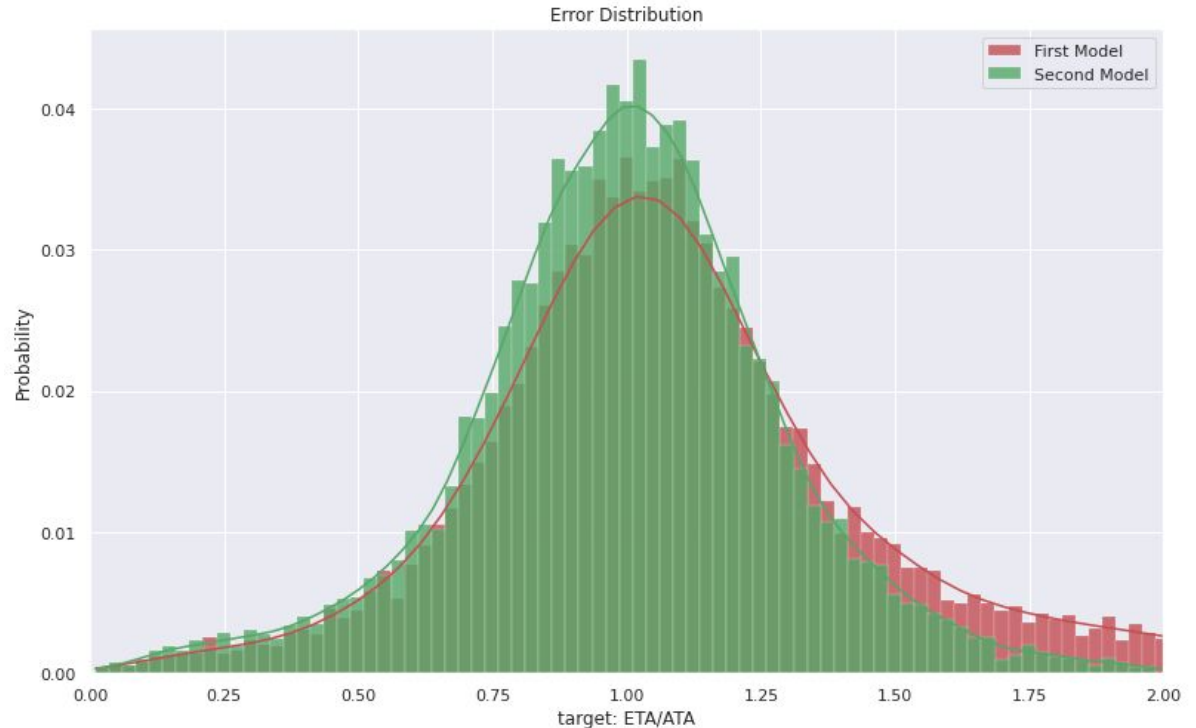
Data Modeling

mean

22.191274

Result

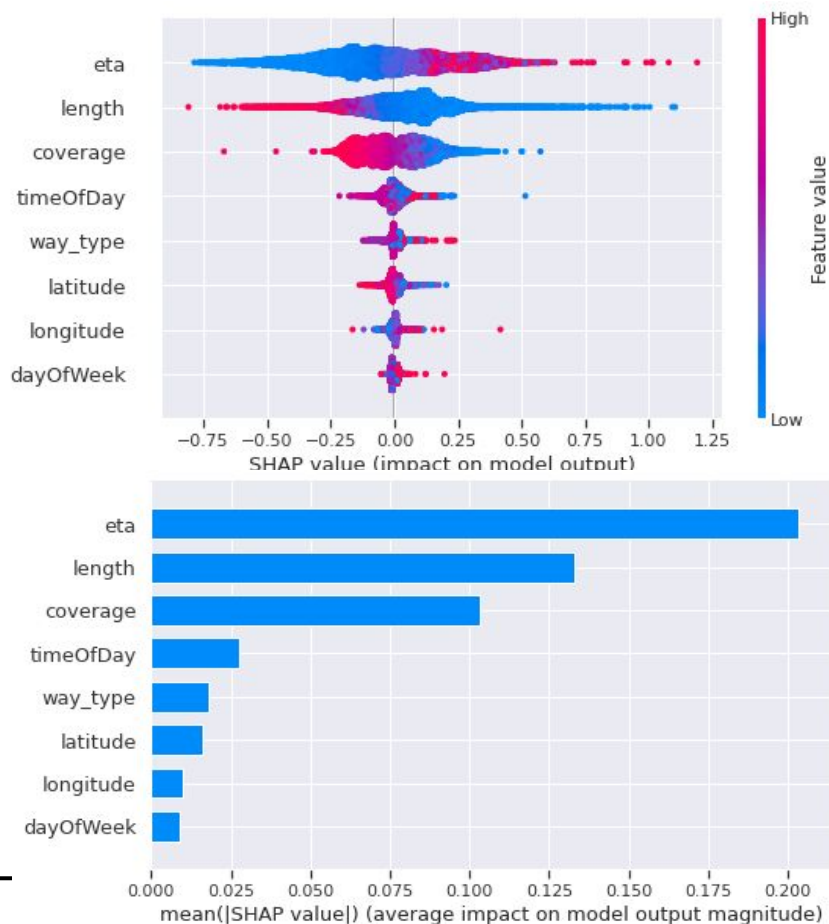
- MAPE: **22.2%**
- First model is the eta of raw data, and second model is my estimator.
- The model reduced the overestimated cases and average is reduced.



Data Modeling

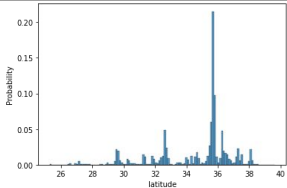
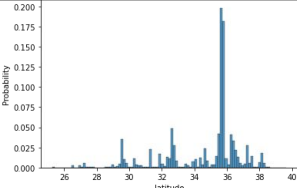
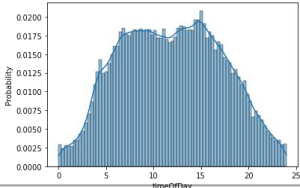
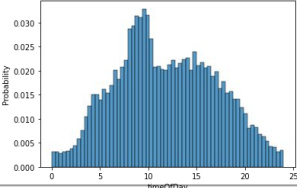
Model Explanation

- I have scored the features by the SAHP value.
- As you can see, the most important features are **ETA**, **Length** and **Coverage**.
- High value of length and low value of eta decrease the output fraction by the model.
- Importance of the features are confirming our first insight from data.
- The least important feature is DayOfWeek.



Challenge

Do the train and test come from the same distribution? I assumed that the distribution is the same, but according to data it is not.

Feature	Train	Test	Notes
isTunnel	0: 99.8%	0: 99.9%	Same
Location			Almost the same
TimeOfDay			Not an exact match
wayType	[3,6,7,9]: 98.2%	[3,6,7,9]: 43.5%	Different

As we only have about 900 samples out of 150'000 samples in train with **waytype** 8 and 10 which is about 54% in test data, we can over/sub sample or evaluate the model by new metric (normalized with the test's distribution MAE), but because of the minor impact of waytype on model and not having enough data for applying some methods to balance it, I negotiated working on this feature.
