Forecasting Traffic Congestion in San Francisco Based on Real-Time Data

Abstract

Traffic and road congestion forecasting is an ever important problem in America with the increase in car ownership and need for transportation. Our paper proposes a novel approach to time-series related problems that uses a single generalized Light Gradient Boosting Machine (LightGBM) model to predict traffic congestions across multiple roads in San Francisco. We find that our approach not only reduces the latency of machine learning models during training, but also outperforms conventional time-series models, i.e. SARIMAX by root mean squared error.

1 Introduction and Background

With the rise of ride-sharing apps and other forms of intelligent transport technology, there is an ever-growing need to predict traffic flow for individuals, public services and businesses alike. Forecasting traffic flow allows businesses to plan meetings and resource transportation in an optimized manner and minimize time wasted; police departments in cities can better allocate their human resources and only deploy its workers in times of predicted congested traffic; individuals rely on Google Maps or Uber traffic time forecasts to plan their travels accordingly. Past academic work, such as Oh et al. (2021) and Zheng and Huang (2020), has used both traditional and "post-AI" time series methods such as ARIMA models and LSTM neural networks. Our paper aims to show that traditional statistical methods such as SARIMAX are restricted in function: they can only predict one road at a time, and do not leverage any 'learning' of trends across roads in an entire city. We aim to show a novel generalized approach that allows us to achieve state-of-the-art prediction across multiple roads with reduced latency, while maintaining usefulness across an even wider range of roads in San Frnacisco.

2 Methods

For this project, we utilize two approaches: a baseline time-series model alongside a novel gradient-boosting decision tree approach.

2.1 Traditional Time Series Models: SARIMAX

We first apply traditional time series models to our forecasting problem. For these models, we consider *undirected* street segments: in other words, the street segment starting at node A and ending at node B is not distinguished from the same street segment starting at node B and ending at node A. In this sense, our baseline only focuses on general traffic density.

We chose the SARIMAX model due to its flexibility (in having many hyperparameters), the presence of seasonality trends in our data, as well as having the option to include hand-crafted exogenous features in our regressions. Indeed, consider a single segment of road, OpenStreetMap ID#195604802 (Shattuck Way, Oakland, CA), which has the most entries (≈ 1950) in our pre-processed dataset, corresponding to about $81 \approx \frac{1950}{24}$ days worth of data. Computing the autocorrelation function (ACF) of this filtered dataset gives:

Preprint. Under review.

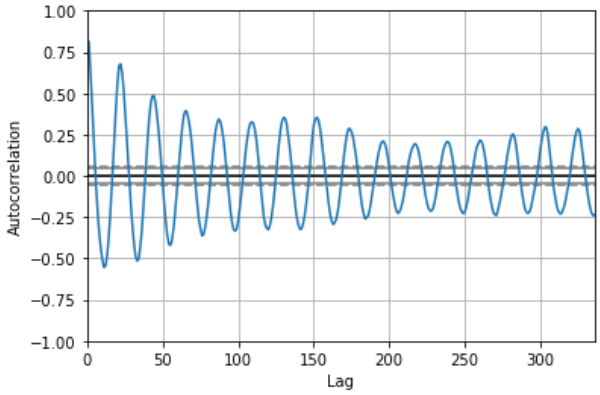

Figure 1: ACF for Time Series of Speeds Across Way #195604802, October-December 2019

As expected, there is a strong positive correlation at lag $24n$ hours and a strong negative correlation at lag $24n - 12$ hours for each integer $n \geq 1$, which gets weaker as n increases. Therefore it makes sense to include a possible seasonality component in the regression.

Recall that we say a process $\{x_t\}$ is SARIMA$(p, d, q) \times (P, D, Q)_s$ if we can write

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t, \tag{1}$$

where:

- B is the backshift operator $Bx_t = x_{t-1}$ (and B^s is the s-times backshift operator $B^s x_t = x_{t-s}$),
- ∇ is the differencing operator $\nabla x_t = (1 - B)x_t = x_t - x_{t-1}$ (and ∇^d is the d-times differencing operator $(1 - B)^d$),
- $\nabla_s^D = (1 - B^s)^D$,
- Φ_P, ϕ, Θ_Q, and θ are polynomials with constant term 1 and degree P, p, Q, and q respectively,
- δ is an intercept term,
- w_t is a Gaussian white noise process.

To define the SARIMAX model with parameters $(p, d, q) \times (P, D, Q)_s$, where we have *exogenous inputs* $r_t^{(j)}$ for $1 \leq j \leq m$ (depending only on time t), we simply add these inputs with associated parameters to the left-hand side of 1:

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t + \sum_{j=1}^{m} \beta_t^{(j)} r_t^{(j)} = \delta + \Theta_Q(B^s)\theta(B)w_t. \tag{2}$$

Here, j keeps track of the "category" of exogenous input being fitted.

As is implied by their names, the SARIMA and SARIMAX models are useful for time series that are non-stationary due to seasonality observations (and note that when $s = 0$, they become ARIMA and ARIMAX models, respectively). After fixing the hyperparameters p, d, q, P, D, Q, s, b, the coefficients of the polynomials Φ_P, ϕ, Θ_Q, θ as well as the $\beta_t^{(j)}$ can be estimated using MLE.

To find the optimal hyperparameters, we used a grid search. In other words, we fitted SARIMAX$(p, d, q) \times (P, D, Q)_s$ models to the data of road segment #195604802 for each of $1 \leq p \leq 3, 0 \leq d \leq 1, 0 \leq q \leq 2, 0 \leq D \leq 1, s \in \{0, 12, 24\}$, both with and without exogenous features. These exogenous features were:

- The standard deviation of the speeds along a given path during a given hour.
- A binary variable indicating whether the data point was taken during a rush hour (7am-9am; 5pm-6pm).

- A binary variable indicating whether the data point was taken on a weekend day.
- The hour number of the data point.

We also restricted to the case $P = Q = 0$ for simplicity and ease of fitting. This means that the seasonality s can only appear in the ∇_s^D term.

2.2 Generalized Gradient Boosting Decision Tree (GBDT) Approach

While SARIMAX is a robust model across trend-related datasets, there are multiple weaknesses that make it ill-advised for the task of forecasting traffic congestion. First, SARIMAX can only be modelled for one time series, while a city comprises of thousands of roads; trends observed across a district of roads or across the city in general cannot be generalized to predict congestion across other nearby roads. Furthermore, the assumption of 'seasonal' data may not be fulfilled. Although the average traffic speed from 24 hours ago may be most relevant for making a road speed prediction given the autocorrelation function, missing timestamps across the night may result in this information being presented at an earlier point in the test set than the SARIMAX model was trained to do.

In this project, we propose a generalized framework for time-series problems that takes in certain exogenous features of a specific road at the desired time-point t for prediction, alongside a history of the road's features across 96 hours, all as an input vector x. As we will show in the results section, this approach yields us promising results compared to the baseline SARIMAX model as introduced in the previous subsection.

2.2.1 Model Selection

While we ran preliminary baseline models with Linear Regression and attained comparable results to SARIMAX, our final results were obtained through training a non-linear model and, more specifically, a gradient boosting decision tree (GBDT).

Gradient Boosting Decision Trees are a popular algorithm class used in classification and regression time-series problems and has yielded strong results across financial applications as well. It refers to a methodology where, at each step, a new decision tree is trained to predict the residual error of the previous tree; the final prediction model is then the ensemble of these decision trees.

In other words, consider a gradient boosting algorithm across M stages, and consider any stage $1 \leq m < M$. Suppose at this stage m suppose there exists an imperfect model F_m for the task. For stage $m + 1$, we train a hidden estimator $h_m(x)$ on the residuals $y - F_m$, such that we minimize the error of

$$h_m(x_i) = y_i - F_m(x_i)$$

across all data points x_i. This yields us an ensemble model

$$F_{m+1}(x_i) = F_m(x_i) + h_m(x_i) = y_i$$

which obtains lower error across the training set.

Two of the most widely known GBDT models across academia and industry are **eXtreme Gradient Boosting** (XGBoost) Chen et al. (2016) and **Light Gradient-Boosted Machine** (LightGBM) Ke et al. (2017). The major difference between the two algorithms is that trees grow depth-wise in XGBoost while in LightGBM, trees grow leaf-wise. While we experiment with both models for the purposes of this project, LightGBMs have historically performed better due to the higher loss reduction and greater model flexibility as a result of a vertical (leaf-wise) growth approach.

3 Related Work

Our overall methodology draws the most inspiration from Zheng and Huang (2020), which compares the performance of a long short-term (LSTM) with a standard backpropagation neural network (BPNN) and an ARIMA model in forecasting traffic flow across a single road. They found that LSTM outperformed BPNN, which outperformed ARIMA, in terms of the RMSE error metric. In particular, the ARIMA model, and the BPNN to a lesser degree, had difficulty predicting the transition from periods of low traffic to high traffic and vice versa (cf. Figures 5-10, Zheng and Huang (2020)), such as during rush hour periods. More specifically, the LSTM was the only model able to accurately

predict these volatile periods in “real-time"; the BPNN and ARIMA models had varying amounts of time-lag when forecasting these periods, leading to the higher RMSE.

4 Data

We use the “Street Speeds" dataset for San Francisco, obtained from `https://movement.uber.com/?lang=en-US` for the months of October to December 2019, which gives aggregate data of Uber trips in San Francisco in a given month. Each data point consists of a timestamp (month/day/year and hour from 0 to 23), identifiers for a single street segment/way (along with identifiers for the beginning and ending junctions), and the average/standard deviation speed of Uber trips using that street segment in the indicated hour; for the generalized model, a data point also consists of this information for each of the last 96 recorded time steps. A segment is listed in the dataset for a given day and hour only if there were more than 5 trips along that segment during that time period. To avoid duplicate entries in the dataset, we perform preprocessing by averaging out all speed information for any given day, hour, and street segment.

4.1 Training Set

For both the SARIMAX and the novel generalized model, we begin by training the model solely across training data across the same 20 busiest roads across the months October to December of 2019. This dataset consists of 24,811 data points across the twenty roads and three months, resulting in an average of 1240.6 time points ($\sim$51.69 days) per road.

For our GBDT model, we experimented with expanding the dataset to other seen roads across San Francisco and testing our model's ability to extrapolate general trends in traffic data across San Francisco. We constructed a dataset of 304,714 data points across 250 of the busiest roads of San Francisco, with no point in our dataset coming from a time point later than the earliest time seen in our test set.

4.2 Test Set

We aim to evaluate the model across 20 of the busiest roads across the months October to December of 2019. Our test set comprises of a total of 6,200 data points across the twenty roads and three months, resulting in an average of around 310 timepoints ($\sim$12.9167 days) per road; this test set remains unchanged for both the SARIMAX and the generalized GBDT model.

5 Results

5.1 SARIMAX

From our grid search for the most popular road in San Francisco, we found that the optimal hyperparameters (with respect to root-mean-square error) in this range were $(p, d, q) = (2, 0, 1)$ and $D = s = 0$, with an RMSE of 1.828 (without the above exogeneous features) and 1.795 (with the above exogeneous features).

To further investigate the behavior of our model, we fit it to the data of the 20 most common road segments appearing in our dataset. Below are the regression statistics for this ARIMAX$(2, 0, 1)$ model on road #23878997 (South Van Ness Avenue, San Francisco, CA), with exogenous features included. This fit had the best performance, in terms of RMSE, out of these 20 roads.

As we can see from Figure 3, the residuals of this fit are generally small, and qualitatively seem to behave like white noise.

On the 20 roads we considered for the SARIMAX models, the ARIMAX(2,0,1) model had an average RMSE of 2.314, with a standard deviation of 0.98. However, a major drawback of the ARIMAX model is its significantly poorer performance on roads with (even slightly) less data points. If we take the 10 most common roads in our dataset, the average RMSE is 1.957, with a standard deviation of 0.415. However, for the next 10 most common roads, the RMSE and standard deviation both jump to

SARIMAX Results

Dep. Variable:	y	No. Observations:	2044
Model:	ARIMA(2, 0, 1)	Log Likelihood	-3377.052
Date:	Tue, 13 Dec 2022	AIC	6772.104
Time:	13:49:53	BIC	6822.708
Sample:	0	HQIC	6790.665
	- 2044		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	23.5722	0.148	158.782	0.000	23.281	23.863
x1	-0.0512	0.009	-5.865	0.000	-0.068	-0.034
x2	-0.2388	0.015	-15.722	0.000	-0.269	-0.209
x3	0.2519	0.164	1.538	0.124	-0.069	0.573
x4	-0.5942	0.093	-6.388	0.000	-0.777	-0.412
ar.L1	0.5876	0.101	5.827	0.000	0.390	0.785
ar.L2	0.1095	0.047	2.336	0.019	0.018	0.201
ma.L1	-0.2171	0.102	-2.136	0.033	-0.416	-0.018
sigma2	1.5940	0.021	75.984	0.000	1.553	1.635

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	12706.30
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.20	Skew:	-0.06
Prob(H) (two-sided):	0.02	Kurtosis:	15.21

Figure 2: Regression Statistics for the ARIMAX$(2, 0, 1)$ Model on Speeds Across Way #23878997

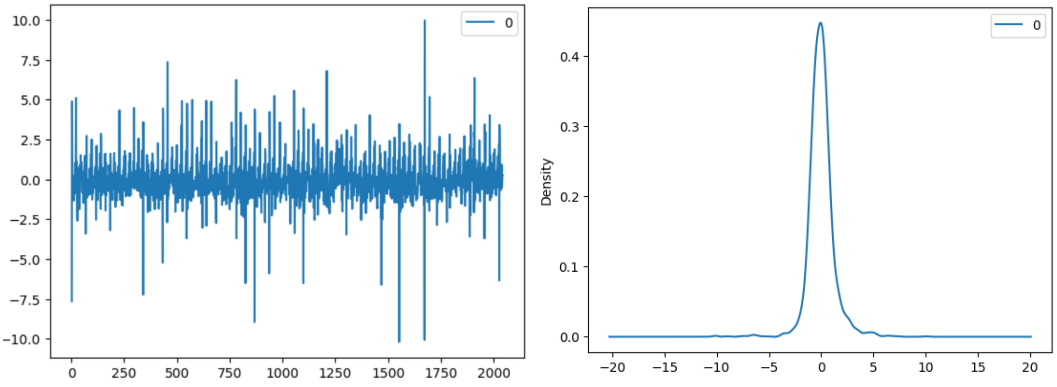

Figure 3: Residuals (left) and residual density (right) for the fitted ARIMAX(2,0,1) Model on Way #195604802

2.671 and 1.254 respectively, which indicates the ARIMAX model's precision issues with respect to smaller number of data points.

It is also interesting to note that a SARIMAX$(2, 0, 1) \times (0, 1, 0)_{24}$ model (i.e. with 24-hour differencing) does *worse*, on average, than the non-differenced ARIMAX model. Indeed, just looking at the 10 most common roads, this SARIMAX model has an average RMSE of 2.048 and a standard deviation of 0.459, and moreover the RMSEs are worse when comparing the models on each individual road. This may be due to an issue with parameter overfitting—for instance, since exogenous features corresponding to the hour of a data point are included in our model, the differencing may be overemphasizing this 24-hour cyclicity.

5.2 Generalized GBDT Model

Across the same dataset of just the twenty most popular roads across San Francisco, our generalized model approach has yielded much stronger results on the test set, with a RMSE of 2.190 by our initial Linear Regression model compared to a RMSE of 2.314 across twenty SARIMAX models for twenty different roads. Furthermore, we were able to obtain better test errors when using LightGBM and XGBoost, with RMSEs of 2.008 and 2.130 on the test dataset respectively; a breakdown of the RMSE errors for all twenty roads per model can be seen in Tables 1 and 2. This shows that our approach of using a generalized model is a strong alternative to simply running a time-series model one road at a time.

Running a grid search, we experiment with the different hyperparameters of the LightGBM as well as the training data size. Sampling from too many roads may result in learning specific patterns from roads that are odd and not common found across roads in the city, whereas sampling from too little roads results in little extrapolation and barely any progress from the baseline model presented in the previous paragraph. We find that with a dataset of 100,000 time points across the 250 aforementioned roads, a LightGBM model of learning rate 0.05, number of leaves set to 275, feature fraction of each tree as 0.9, bagging fraction as 0.8 and bagging frequency at 10 yields us a model with a RMSE of 1.983 across all twenty roads and 1.791 across the first ten roads, with a similar standard deviation of 0.989. Across the twenty most popular roads, our generalized LightGBM model obtained a lower RMSE across nineteen roads despite not being specifically trained for a singular road, while taking comparatively minimal time to train and fine-tune. That being said, across all iterations of our generalized GBDT algorithm, we observe that our model is still prone to underestimating the change in speed at the rush hours of the road.

	Root Mean Squared Error			
Road ID	SARIMAX	LightGBM	XGBoost	Linear Regression
195604802	1.795	1.475	1.543	1.648
215346291	2.079	1.765	1.863	1.901
7700816	2.661	2.556	2.804	3.160
22372749	2.134	1.821	1.933	1.897
6331826	1.828	2.243	2.377	2.378
23878997	1.272	1.302	1.336	1.488
28436962	1.958	1.783	1.891	1.853
125126193	2.096	1.735	1.897	1.800
84733302	1.393	1.009	1.112	1.206
7699312	2.357	2.474	2.686	2.655
575440218	2.697	2.471	2.608	2.713
184573421	5.891	5.548	5.846	6.044
205501690	2.267	2.056	2.171	2.164
449859227	2.074	1.633	1.691	1.700
203305863	2.459	2.144	2.425	2.568
225806029	1.474	1.291	1.412	1.676
24501620	2.783	2.295	2.338	2.495
443319402	3.346	2.976	3.088	3.020
191010941	1.921	1.659	1.733	1.769
201989365	1.797	1.249	1.378	1.384
Average RMSE:	**2.314**	**2.008**	**2.130**	**2.190**

Table 1: A breakdown of the root mean squared error across twenty of the most popular roads in San Francisco across four different algorithms: the SARIMAX baseline, then the generalized model trained via a LightGBM, XGBoost, and a Linear Regression Model. All four algorithms are trained on only data from the twenty most popular roads.

Road ID	RMSE of SARIMAX	RMSE of Best LightGBM	RMSE Improvement
195604802	1.795	1.519	0.276
215346291	2.079	1.776	0.303
7700816	2.661	2.579	0.082
22372749	2.134	1.861	0.274
6331826	1.828	2.225	-0.397
23878997	1.272	1.229	0.044
28436962	1.958	1.732	0.226
125126193	2.096	1.663	0.433
84733302	1.393	0.997	0.396
7699312	2.357	2.333	0.024
575440218	2.697	2.428	0.270
184573421	5.891	5.686	0.205
205501690	2.267	2.027	0.240
449859227	2.074	1.609	0.465
203305863	2.459	2.188	0.271
225806029	1.474	1.307	0.167
24501620	2.783	2.167	0.616
443319402	3.346	2.931	0.415
191010941	1.921	1.602	0.318
201989365	1.797	1.247	0.550
Average RMSE:	**2.314**	**1.983**	**0.331**

Table 2: A breakdown of the root mean squared error across twenty of the most popular roads in San Francisco for our best LightGBM model, compared to the baseline SARIMAX model. A RMSE improvement is observed across nineteen of the twenty roads.

5.3 Feature Analysis

Aside from analyzing the performance of our generalized GBDT model on the Uber Movement dataset, we also investigate the importance of different features in explaining the model. To analyze the relative importance of a single feature in the entire dataset, we calculate Shapley scores, which aims to compute the average expected contribution of each feature on the prediction outcomes. On average, our model determined that by far the five most important features are the average speeds of the past two time lags, followed by the current hour from 0-23, and then the information from the 20th and 21st previous time lag. Overall, this seems to corroborate with our previous belief that there are less than 24 timesteps of information recorded per day, and that the 20th and 21st previous time lag is most likely giving information regarding the previous day at a similar time. A bar plot of the twenty most important features can be seen in Figure 4.

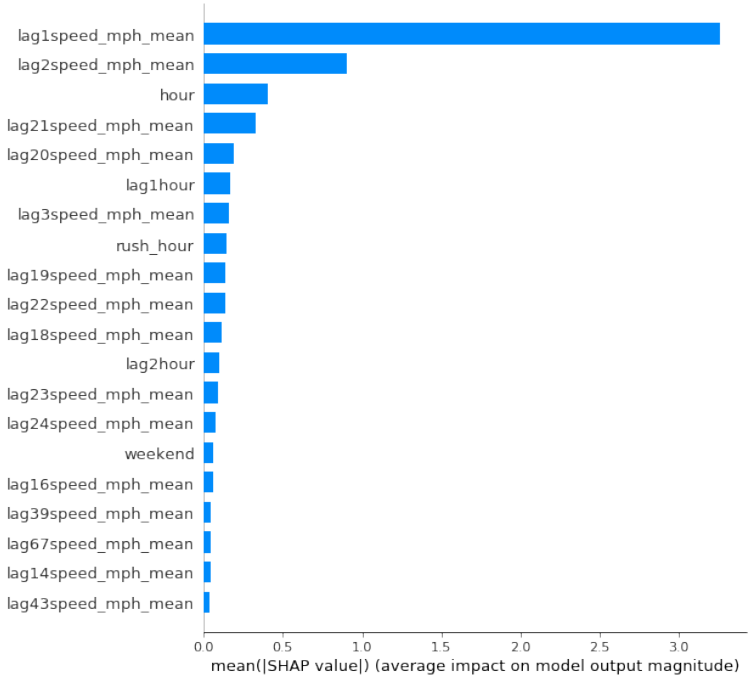

Figure 4: Shap Values of the twenty most features for our final LightGBM model. Note that information from 16-24 hours ago are extremely important, as is information from the previous two hours. Our constructed exogenous features weekend and rush hour seem to also be among the more important features as well.

6 Conclusion

Our paper proposes a novel approach to time-series related problems that not only outperforms SARIMAX by root mean squared error, but also reduces the time taken to train a model by constructing a generalized framework across all roads in San Francisco. Furthermore, we have shown that the exogenous features of determining rush hour and weekend information has yielded stronger results across all model experiments. We believe that our contributions in the field of movement speed predictions will not only be applicable across the industry, i.e. using information from San Francisco or greater American roads to predict roads in large cities across Canada or the United Kingdom, and across time-series decisions, i.e. being able to predict equity stock futures across a multitude of industries using a single model. Future work will look to see if there are other ways to further improve upon the results of LightGBM, and rather any sort of isotonic regression or calibration from predicted to actual values will prevent the model from systematically underestimating road speed at rush hours.

References

YongKyung Oh, JiIn Kwak, JuYeong Lee, and Sungil Kim, Time delay estimation of traffic congestion propagation based on transfer entropy, (2021).

Jianhu Zheng and Mingfang Huang, Traffic flow forecast through time series analysis based on deep learning, *IEEE Access 8* (2020), 82562–82570.

Tianqi Chen and Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, (2016), *Association for Computing Machinery* (2016), 10.1145/2939672.2939785.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, LightGBM: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems) (2017), Pages 3146-3154*