Predicting BC Ferry Delays

Christopher Wang, Alex Trostanovsky, Andrea Pagotto

Introduction

. Given historical BC weather, traffic, and ferry-schedule data, we set out to learn a model that can accurately predict ferry departures that will be delayed. The model used for our final score in Kaggle was an ensemble of 4 models trained on different subsets of features, resulting in 6th place overall.

Figure 1. BC Ferry Routes

Data Overview

We analyzed a data set with 49,500 ferry trips from 2016 and 2017, to predict whether ferries will be delayed or not for 12,400 trips between November 2017 and March 2018..

Data Insights:

- The delayed class is 18.1% of the total data
- Ferries and routes are highly related (Fig. 3)
- Number of delays has seasonal patterns (Fig. 4)
- Delays are primarily traffic related
- Traffic strongly affects particular routes and ports (Horseshoe Bay)

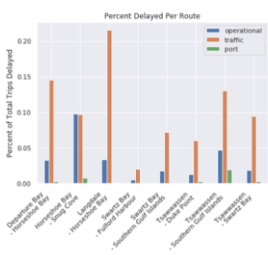

Figure 2. Ferries vs Routes

Figure 3. Ferries vs Routes

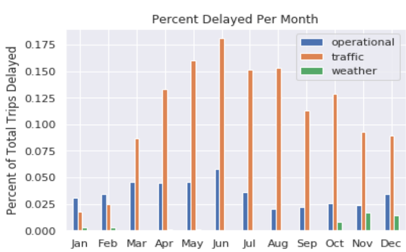

Figure 4. Monthly delays

Methodology

Table 1. Feature Engineering

	Given Features	Generated Features
Date and Time	Date of Departure Time	Day of year, day of week days of month, month of year, weak of year
Weather	Wind speed, wind direction, visibility, humidity, temperature, dew point	Rolling window min, max, standard deviation, mean, on 5 minute, 15 minute, 30 minute, 1h and 1 day intervals
Traffic	Traffic by minute	Rolling window min, max, standard deviation, mean, on 5 minute, 15 minute, 30 minute, 1h and 1 day intervals
Historical Ferry Trips	Vessel Name Route Name	Statistical features
Holiday	Date of Departure	Holiday Indicator, Days until, Days since, number of trips until and number of trips since
External Ferry Information	Vessel Name Route Name	Ferry age, ferry length, car capacity, passenger capacity, ferry speed, horsepower

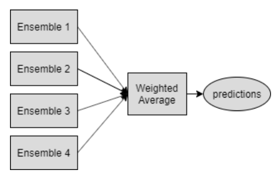

Figure 5. Final Ensemble Diagram

Our approach involves ensembles based on many weak learners, for example Ensemble 1 shown in figure 6. Also several high performing models were ensembled, as shown in figure 5.

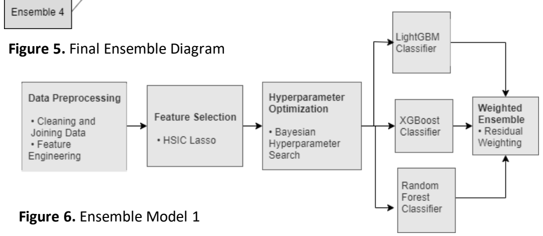

Figure 6. Ensemble Model 1

Results

The following table displays our results from our best ensembles individually, and also their combined best score. The features and classifiers used in each ensemble is also shown. Each ensemble had varying features, classifiers and hyperparameter tuning.

Table 2. Ensemble Model AUC-ROC scores

Model	Features	Classifiers	Public Leaderboard	Private Leaderboard
Ensemble 1 (50%)	date, time, holidays, traffic features	LightGBM[1], XGBoost[2], Random Forest	0.71157	0.71804
Ensemble 2 (12.5%)	traffic features, weather	LightGBM[1]	0.69970	0.70620
Ensemble 3 (12.5%)	ferry and route generated features	XGBoost[2]	0.70530	0.70127
Ensemble 4 (25%)	Port specific traffic features	LightGBM[1], XGBoost[2], Random Forest	0.71566	0.70477
Final Ensemble	All of the above	(models 1, 2, 3, 4)	0.72740	0.72251

Discussion

Ensemble 1 was the highest weighted model, and the feature importances for this model from SHAP analysis is shown in figure 7. Figure 8 shows a comparison between models used to build a larger model, resulting in a boost in performance.

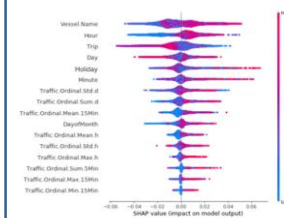

Figure 7: Ensemble 1 feature importances from SHAP analysis

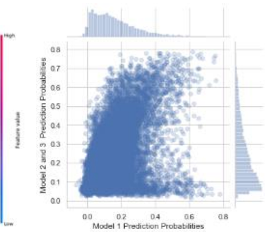

Figure 8: Ensemble 1 vs Ensemble 2 and 3 prediction probabilities, AUC-ROC score was 0.72447

Insights

What we tried:

- Larger ensembles (1024+ model) of minor hyperparameter variations
- Different model architectures: logistic regression, support vector machines, decision trees, convolutional Neural Networks on 2D representations of features as rolling windows

Machine Learning Insights:

- Basic ensembling >> weighted ensembling
- Fewer more expressive features >> many inexpressive features
- Small expressive ensembles >> Large meaningless ensembles
- Hand crafted features performed better than AutoML
- Trees play better with label encoding than one-hot encoding
- Standard cross validation is not an effective measure for public set

Business Insights:

- Traffic delays - traffic most significantly affects delays: prepare for known high-volume traffic times, offer customer incentives to travel on off-peak times, implement more efficient loading strategies
- Other delays: Schedule regular maintenance for the Spirit of British Columbia to prevent abnormally high amounts (~30%) of maintenance related delays, investigate route Horseshoe Bay - Snug Cove to address high operational delays

Ideas for future work:

- Use additional validation strategies
- Loss reweighting based off delay type/representation in test set

Conclusion

We conclude that for the best predictive results, an ensemble of a variety of models with different predictive distributions should be used. From the perspective of business applications based on the ROC-AUC scores, simple models can perform comparably well.

[1]Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.
[2]Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016