The 20th Summer School of Behavioral Modelling

September 17, 2021

Behavioral Models Based on Weak Learners using Multiple Sensors

Muhammad Awais Shafique, PhD

Severo Ochoa Postdoctoral Researcher

Center for Innovation in Transport (CENIT), Barcelona, Spain

Introduction

- When predicting, the least one can do is Random Guessing

- **Weak Learner**
- "A weak learner produces a classifier which is only slightly more accurate than random classification."

Pattern Classification Using Ensemble Methods, pg 21, 2010

- **Weak Classifier**
- A classifier that achieves slightly better than 50% accuracy.

Introduction

- “For binary classification, it is well known that the exact requirement for weak learners is to be better than random guess.”
- “Notice that requiring base learners to be better than random guess is too weak for multi-class problems, yet requiring better than 50% accuracy is too stringent.”

Ensemble Methods, pg 46, 2012

Introduction

- A popular example is Decision Tree.
- Weakness can be controlled by the depth of tree.
- Weakest tree: only one node and binary decision made on only one variable.

- “Because boosting requires a weak learner, almost any technique with tuning parameters can be made into a weak learner. Trees, as it turns out, make an excellent base learner for boosting.”

Applied Predictive Modeling, pg 205, 2013

Introduction

- **Strong Learner**
- A strong learner produces a classifier that achieves arbitrarily good accuracy, better than random guessing.

- For modeling tasks, we aim to develop a strong classifier that makes predictions with good accuracy with high confidence.
- For instance, applying Support Vector Machines directly to the dataset.

Introduction

- In short

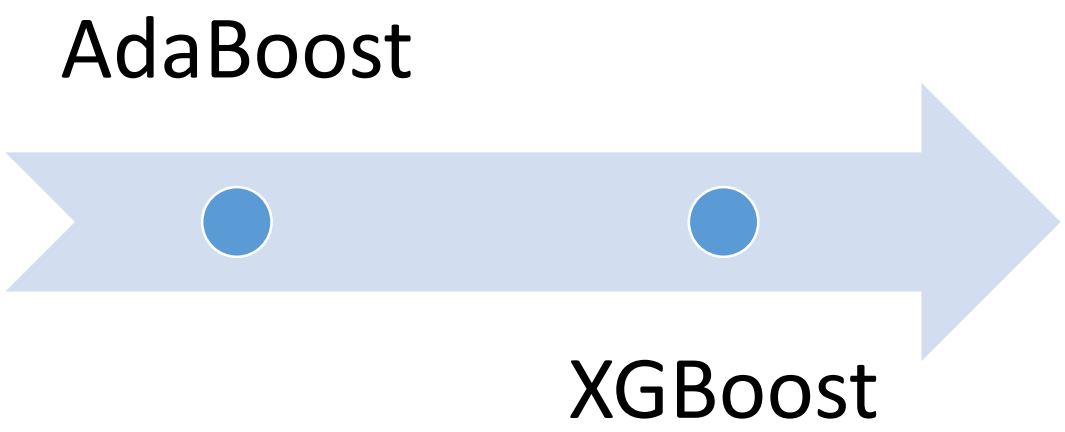
- Weak learners: Slightly better than random.
- Strong learners: Having good or even near-optimal accuracy.

- Are they equivalent?

YES

Boosting

- A strong learner can be constructed from many weak learners.
- This became the basis for boosting methods



Boosting

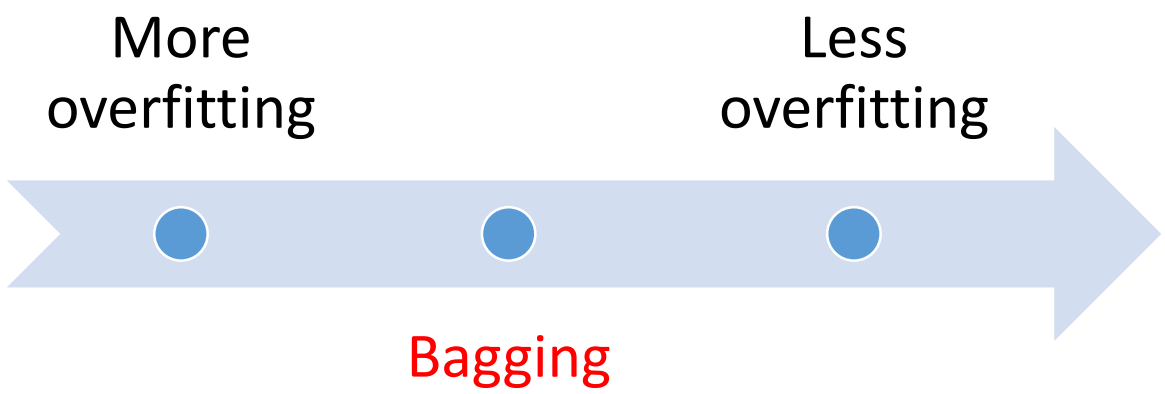
- The goal of boosting ensembles
- Develop a large number of weak learners for a predictive learning problem.
- Combine them in a way to achieve a strong learner.

- Weak learners: Easy to prepare but not desirable.
- Strong learners: Hard to prepare and highly desirable.

Bagging vs. Boosting

Bagging

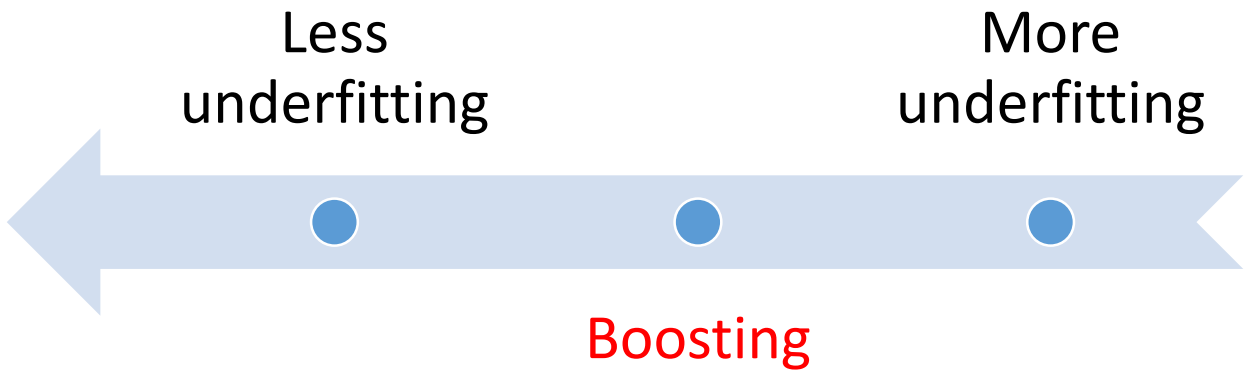
- Train a number (ensemble) of decision trees from bootstrap samples of your training set.
- After the decision trees are trained, we can use them to classify new data via majority rule.



Bagging vs. Boosting

Boosting

- Start with one decision tree stump (weak learner) and “focus” on the samples it got wrong.
- Train another decision tree stump that attempts to get these samples right.
- Repeat until a strong classifier is developed.



Paper for Discussion

Hak Lee, E., Kim, K., Kho, S.Y., Kim, D.K. and Cho, S.H., 2021. **Estimating Express Train Preference of Urban Railway Passengers Based on Extreme Gradient Boosting (XGBoost) using Smart Card Data**. *Transportation Research Record*.

Key Points

- XGBoost vs. MNL
- SMOTE (Synthetic Minority Over-sampling Technique)
- SHAP (SHapley Additive exPlanation)

Introduction

- Express strategy on the urban railway since 2009.
- Line 9-first private subway in Seoul to introduce express trains.
- Local and express trains both on the same railway.
- Local Train – 30 stops (100 min)
- Express Train – 13 stops (60 min)

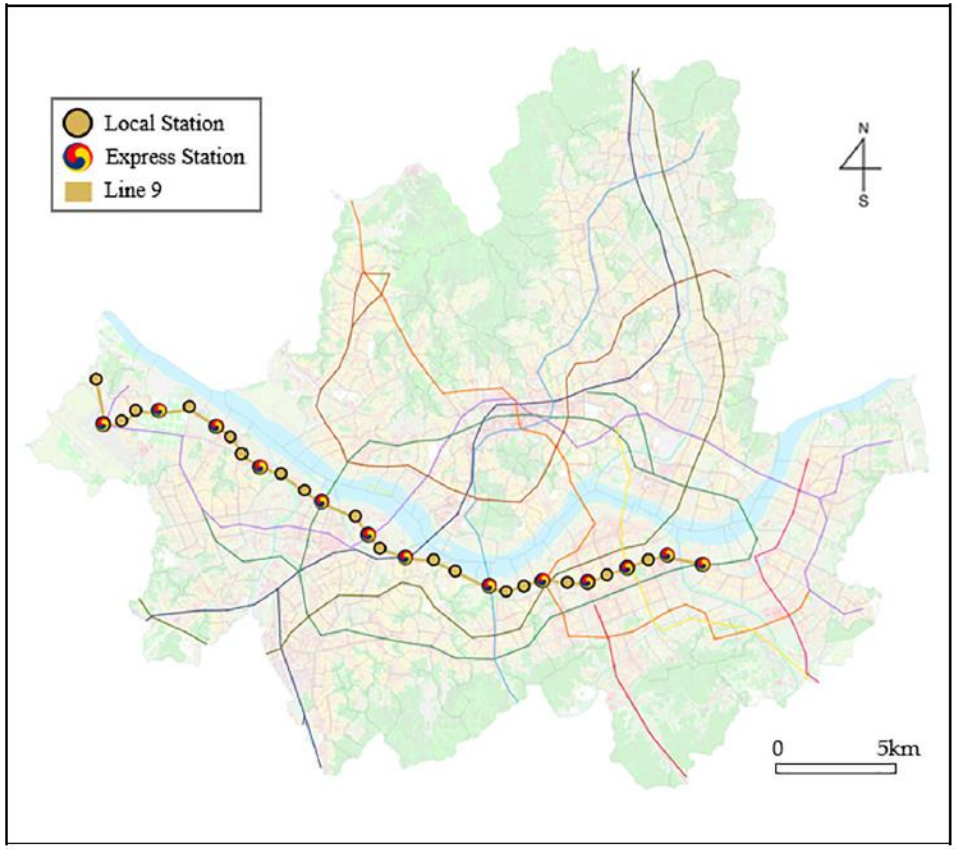


Figure 1. The network of the Seoul Metro Line 9.

Collected Data

Table 1. Description of the Smart Card Data

No.	Detail	No.	Detail
1	Card identification (ID)	14	Total travel distance
2	Transaction ID	15	Total travel time
3	Mode code	16	Boarding fare
4	Line ID	17	Alighting fare
5	Name of the transit line	18	The number of passengers
6	Vehicle ID (for the bus)	19	Boarding violation penalty
7	Vehicle number (for the bus)	20	Alighting violation penalty
8	Boarding station ID	21	Passenger code (general, student, elderly)
9	Alighting station ID	22	Boarding date
10	Name of boarding station	23	Alighting date
11	Name of alighting station	24	Transfer station ID
12	Boarding time	25	Transfer time
13	Alighting time	26	Number of transfers

Collected Data

Table 2. Description of the Train Log Data

No.	Detail	No.	Detail
1	Name of affiliate	5	Train ID
2	Line identification (ID)	6	Train type
3	Arrival time	7	Boarding station ID
4	Direction of train	8	Alighting station ID

Pre-Processing

- One day Smart Card Data and Train Log Data integrated.
- Passenger's boarded train estimated.
- Other train generated as unselected alternative.

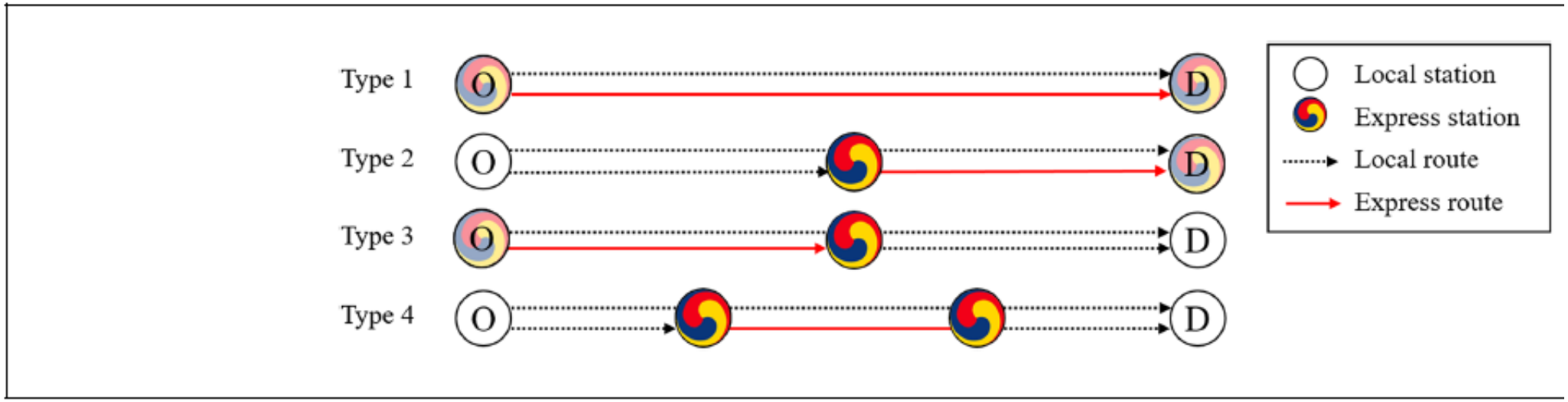
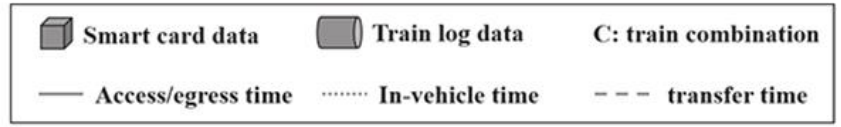
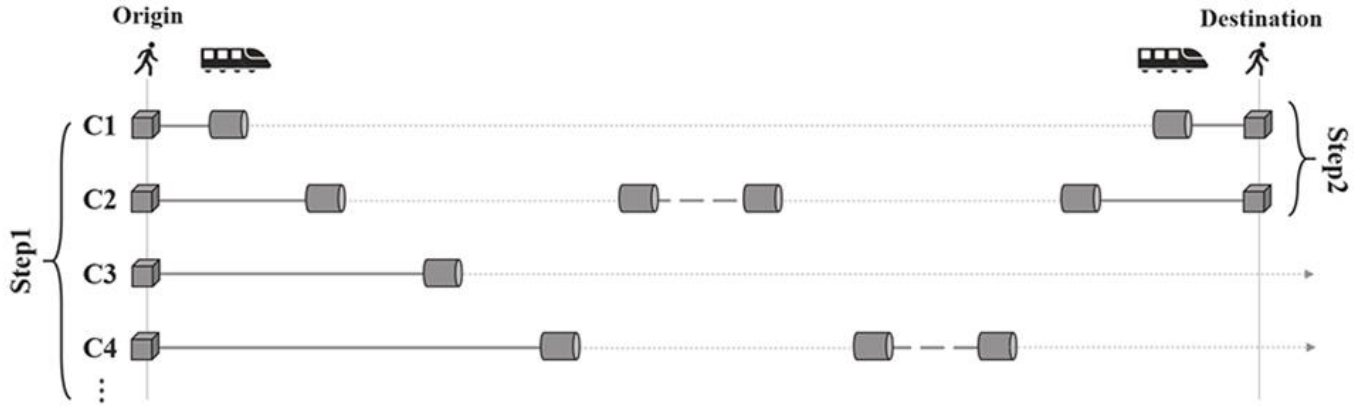


Figure 2. Conceptual diagram of four types of express service.
Note: D = destination; O = origin.

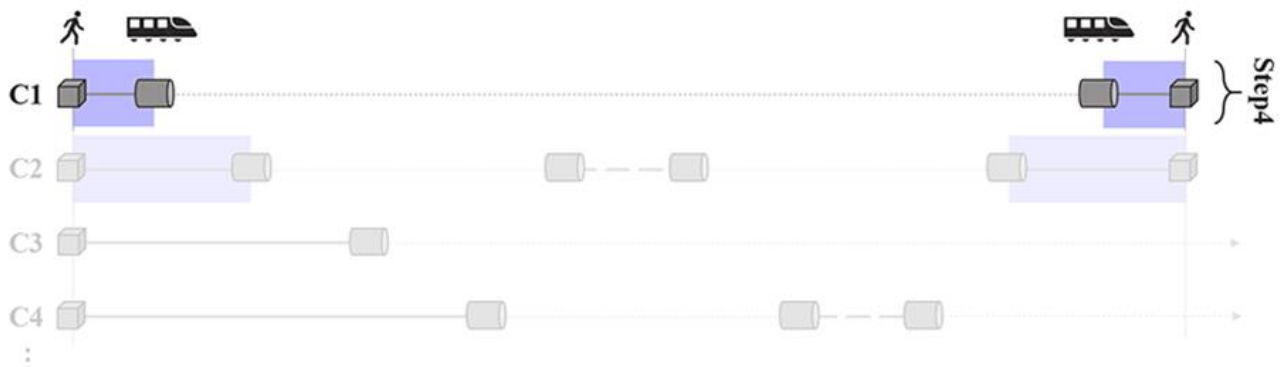


Step1: Generate all train combinations
Step2: Identify the alternative train combinations for each passenger



Step3: Develop CDFs of access/egress time for each station with the passengers who have only one alternative

Step4: Assign the alternative with the highest probability to passengers who have multiple alternatives



Pre-Processing

Table 3. Description of the Pre-processed Data

No.	Detail	No.	Detail
1	Passenger identification (ID)	7	Average total travel time of the local train
2	Type of origin–destination (OD) stations (1–4)	8	Average waiting time of the local train
3	Average total travel time of the express train	9	Average in-vehicle time of the local train
4	Average waiting time of the express train	10	Average crowding of the local train
5	Average in-vehicle time of the express train	11	Number of transfers for the express train
6	Average crowding of the express train	12	Train choice (local: 0, express: 1)

Pre-Processing

- Imbalanced data
- SMOTE (Synthetic Minority Over-sampling Technique) was used.
- SMOTE uses each data point of a minority class and generates new samples along the line joining them to their k-nearest neighbors.
- With XGBoost we don’t need to worry about multicollinearity.
- XGBoost trained on 85% randomly selected data.
- Various hyperparameters were tuned.

Performance Measure

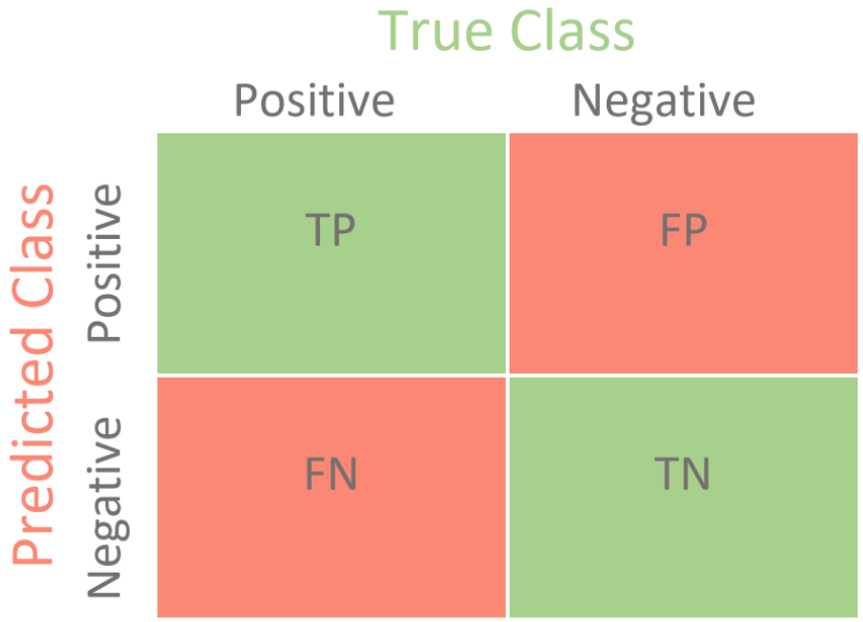
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where
TP is the true positive,
FP is the false positive,
TN is the true negative, and
FN is the false negative.



Performance Measure

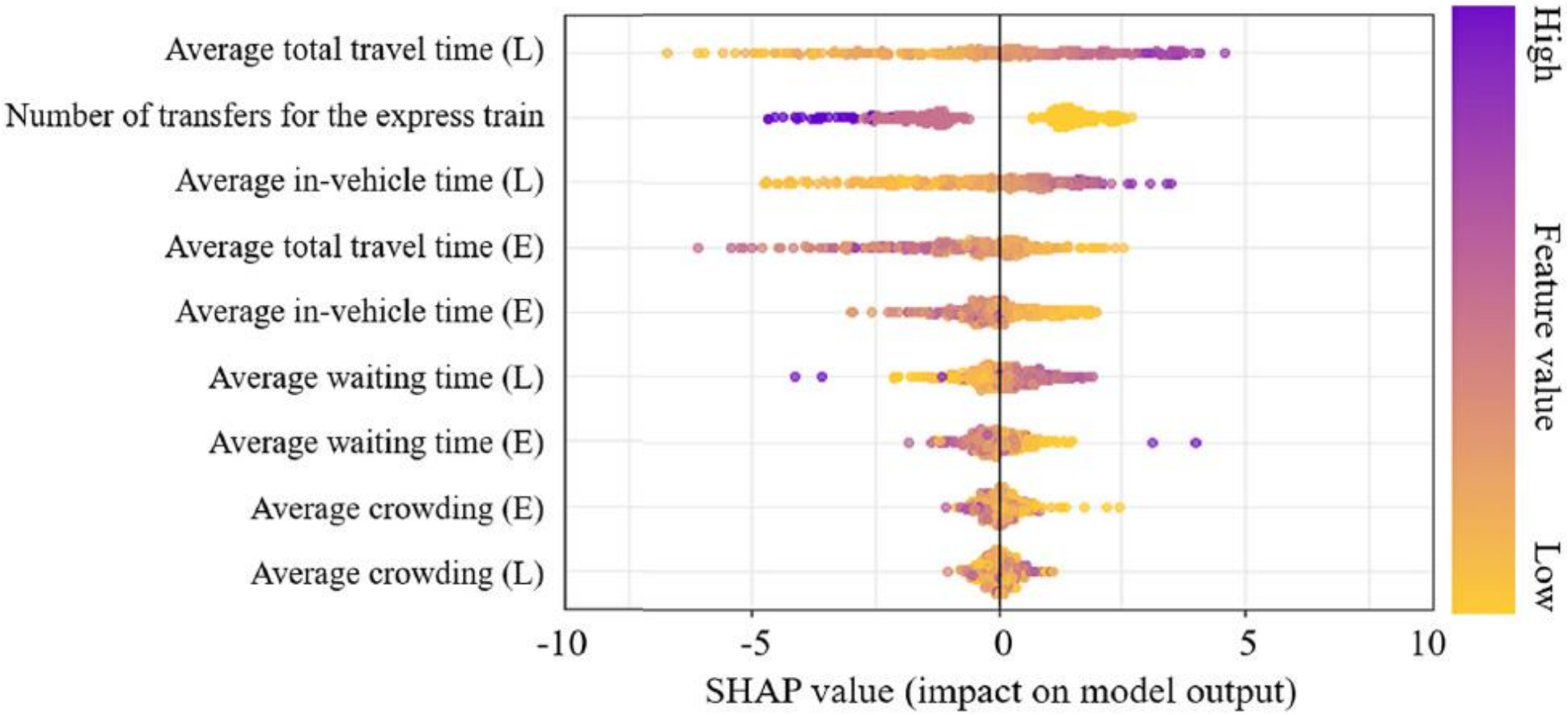
Table 5. Estimation Results of the Express Train Choice with Extreme Gradient Boosting (XGBoost)

Detail	Actual passengers' choice		Model evaluation			
	Local train (trips)	Express train (trips)	Precision	Recall	Accuracy	F1 score
Type 1: No transfer (express train only)	1,410	6,442	0.975	0.997	0.976	0.985
Type 2: One transfer (express-local train)	2,010	1,450	0.959	0.975	0.972	0.967
Type 3: One transfer (local-express train)	2,011	850	0.971	0.995	0.990	0.983
Type 4: Two transfers (local-express-local)	800	151	0.993	0.987	0.997	0.990
Total	6,231	8,893	0.972	0.993	0.979	0.982

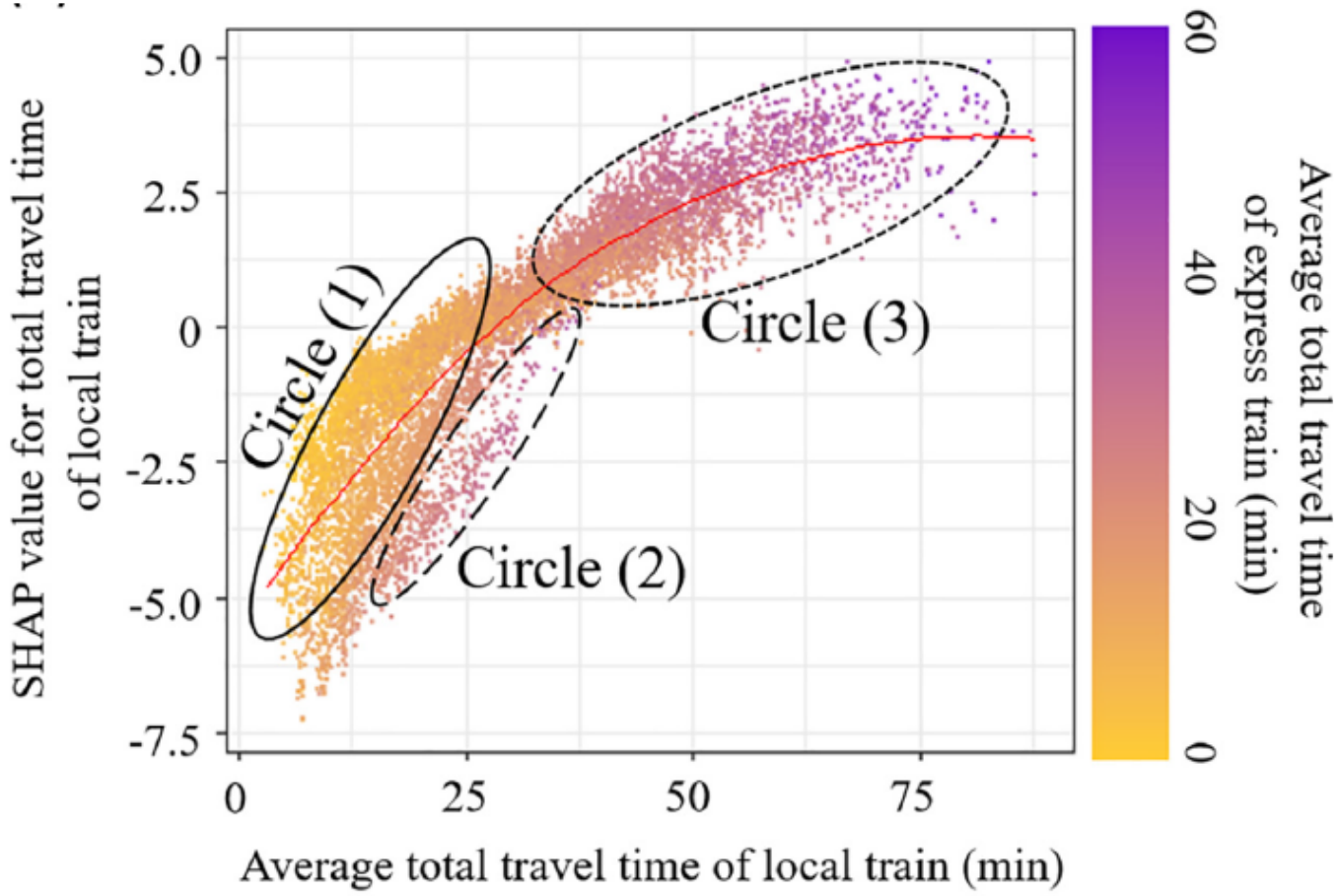
Table 6. Estimation Results of the Express Train Choice with Multinomial Logit (MNL) Model

	Variable				
	Constant	Average waiting time (minutes)	Average in-vehicle time (minutes)	Average crowding (passengers/train)	F1 score
Type 1: No transfer (Express train only)	0.04^{**}	-0.27^{***}	-0.24^{**}	-0.004^{***}	0.937
Type 2: One transfer (express-local train)	-1.70^{***}	-0.27^{***}	-0.24^{**}	-0.004^{***}	0.429
Type 3: One transfer (local-express train)	-2.19^{***}	-0.27^{***}	-0.24^{**}	-0.004^{***}	0.381
Type 4: Two transfers (local-express-local)	-3.03^{***}	-0.27^{***}	-0.24^{**}	-0.004^{***}	0.101
Total					0.720

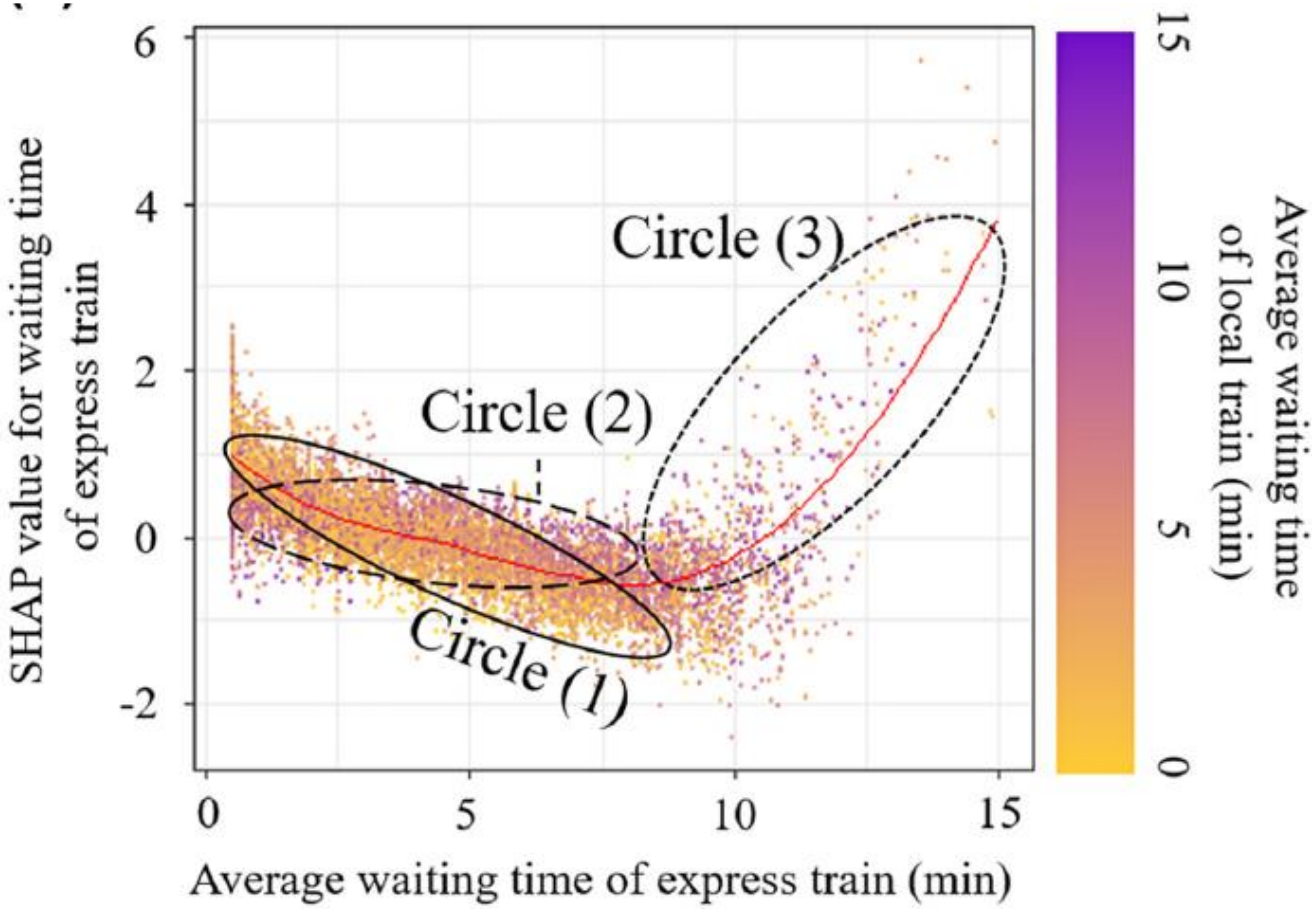
SHapley Additive exPlanation



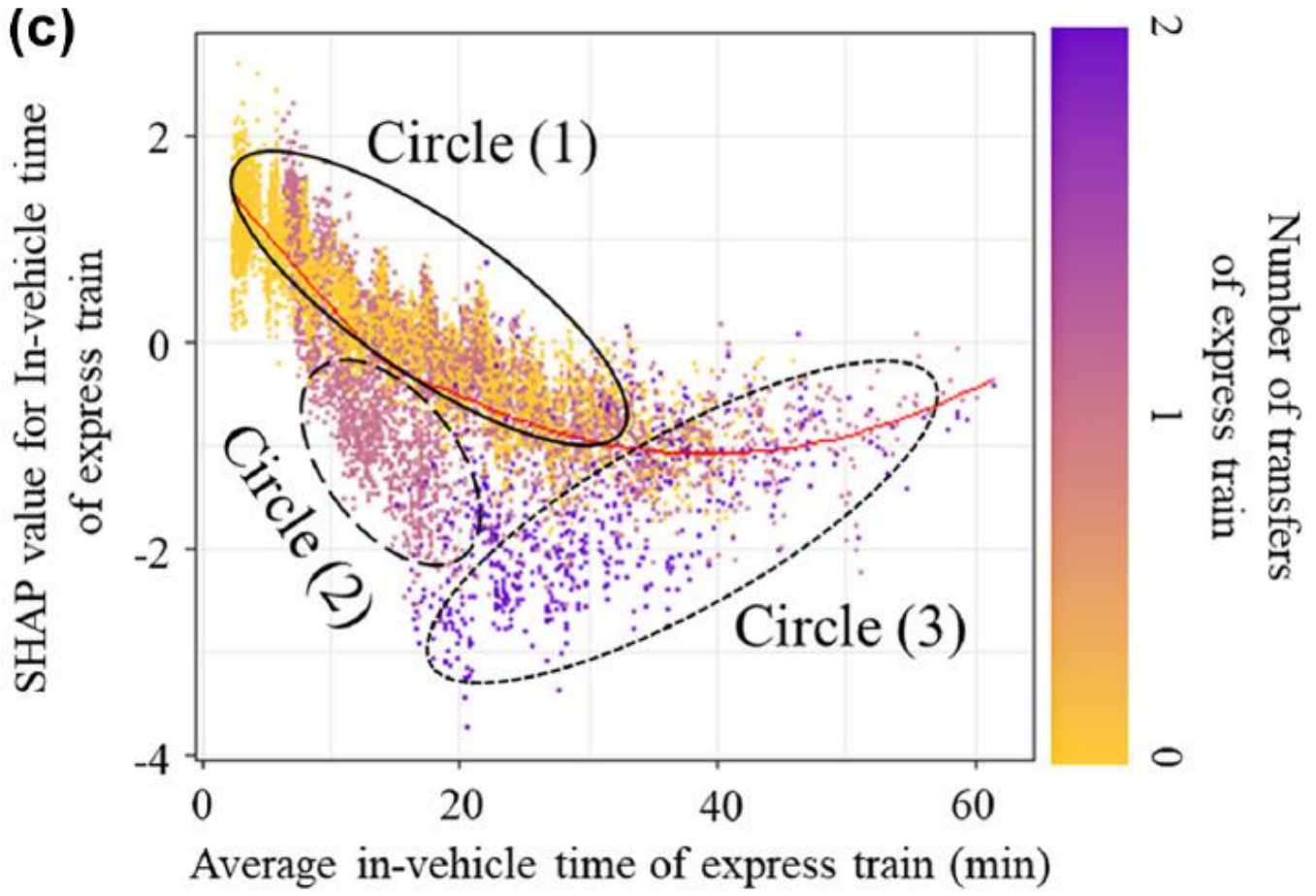
SHapley Additive exPlanation



SHapley Additive exPlanation



SHapley Additive exPlanation



Concluding Remarks

- XGBoost performs better than MNL.
- SMOTE can be conveniently used to address imbalanced data.
- SHAP can be used to understand the impact of each variable.


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
nanni
nandri
dhanyavad
thank you
gracias
хвала
asante
manana
obrigada
tenki
murakoze
bedankt
bayarlalaa
gracie
hvala
mauruuru
koszonom
akun
dankon
aciū
enkosi
djiere dieuf
tau
mochchakkeram
mamnun
дякую
chokrane
dziękuję
chnorakaloutioun
gratias ago
gracies
sulpay
taiku
go raibh maith agat
sobodi
dėkuji
sagolun
sukriya
kop khun krap
obrigado
mesi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
kam sah hamnida
rahmat
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xiexie
merci