Decision-aid Projects

Chen Yifeng, Qing Jun

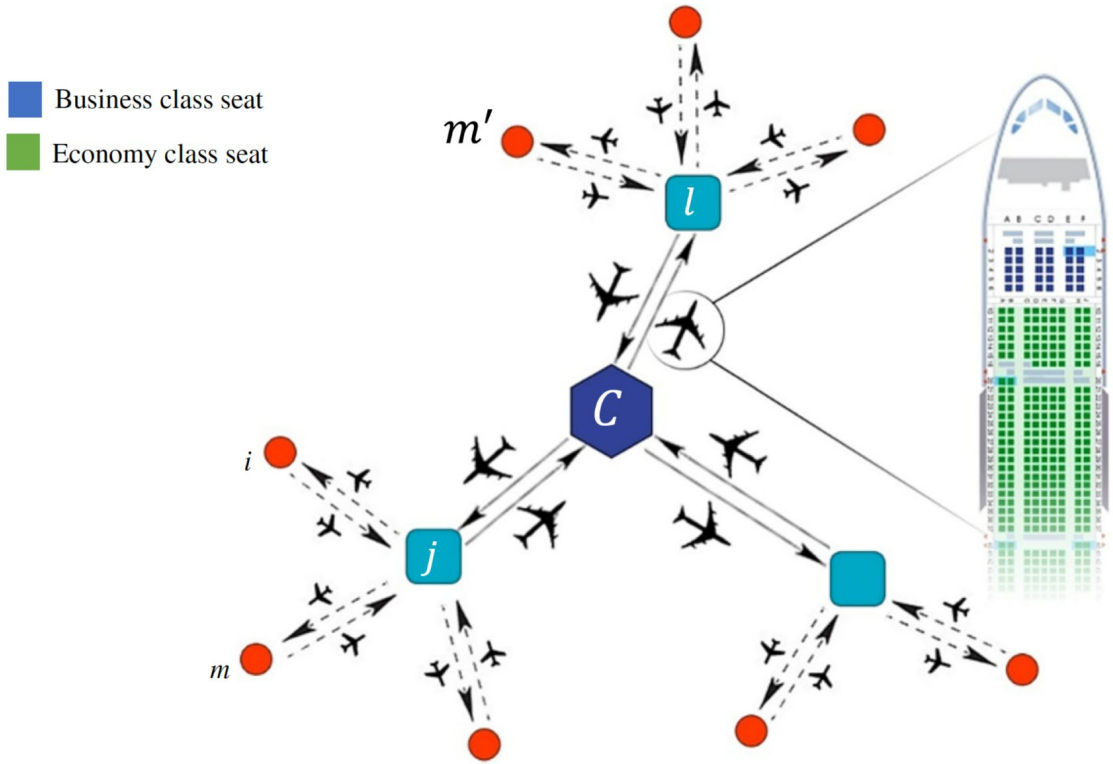
EPFL - 23/06/2022

Project 1

presented by Yifeng

Chen Yifeng, Qing Jun

Background



Deterministic Problem

Purpose:

1. maximize the revenue of selling the tickets to different classes and itineraries.
2. minimise the total transportation costs.
3. minimise the total installation costs.

Only one scenario exists in the deterministic problem.

Input Variables

N	The number of nodes in the hubs-spoke network not considering the central hub.
P	The number of hubs.
K	The number of costumer classes.
$dist_{im}$	Distance from node i to node m.
C_{1k}	unit transfer cost per distance for a customer from class k in a flight between spoke node and hub which is defined as leg type 1.
C_{2k}	unit transfer cost per distance for a customer from class k in a flight between a hub node and central hub which is defined as leg type 2.
d_{imk}	Traffic demand between origin i and destination m for customer class k.
r_{imk}	Ticket price of itinerary between origin i and destination m for customer class k.
$nfNH_{ij}$	Number of flights available for itinerary between node i and hub j (each flight travel from node i to j and in the other direction.
$nfHH_j$	Number of flights available for itinerary between hub j and the central hub (each flight travel in both direction).
Q_1	Number of seats available in a flight at leg type 1.
Q_2	Number of seats available in a flight at leg type 2.
$Fixedcost_j$	Fixed cost for establishing non-central hub j.

Decision Variables

$t_{i,m,k}$ - Int variable. The number of ticket from i to destination j for class k

$x_{i,j}$ - Binary variable. If there is connection from non-hub i to hub j, the value is equal to 1. If node j is selected as a hub, $x_{j,j}$is equal to 1. The value is 0 in other cases.

FNH_i - Int variable. The number of flight needed to take of from non-hub node i.

FHN_i - Int variable. The number of flight needed to arrive to non-hub node i.

ONH_i - Int variable. The number of flight needed to take of from hub i.

OHN_i - Int variable. The number of flight needed to arrive to hub i.

Objective Function

Cost of flight between non-central hub and non-hub

$$
\begin{aligned}
profit &= revenue - cost \\
&= \sum_{i\in I}\sum_{m\in I}\sum_{k\in K} t_{imk} r_{imk} - \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K} C_{1k}\cdot\left(dist_{ij}t_{imk} + dist_{ji}t_{mik}\right) x_{ij} \\
&- \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K} C_{2k}\cdot\left(dist_{j0}t_{imk} + dist_{0j}t_{mik}\right)\left[x_{ij}\left(1 - x_{mj}\right)\right] - \sum_{j\in I} Fixedcost_j x_{jj}
\end{aligned}
$$

Cost of flight between non-central hub and center hub

Cost of building non-central hubs

Constraints

$$\sum_{j \in I} x_{ij} \leq 1 \quad \forall i \in I$$

$$\sum_{j \in I} x_{jj} = P$$

$$x_{ij} \leq x_{jj} \quad \forall i, j \in I$$

Constructing the network

$$t_{imk} \leq d_{imk} \quad \forall i, m \in I, \forall k \in K$$

Demand constraint

$$\lceil \sum_{m \in I} \sum_{k \in K} \frac{t_{imk}}{Q1} \rceil \leq \sum_{j \in I} NFNH_{ij} x_{ij} + M x_{ii}$$

$$\lceil \sum_{m \in I} \sum_{k \in K} \frac{t_{mik}}{Q1} \rceil \leq \sum_{j \in I} NFNH_{ij} x_{ij} + M x_{ii}$$
$$\forall i \in I$$

$$\lceil \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{k \in K} \frac{t_{imk}}{Q2} x_{ij} (1 - x_{mj}) \rceil \leq NFHH_j x_{jj}$$

$$\lceil \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{k \in K} \frac{t_{mik}}{Q2} x_{ij} (1 - x_{mj}) \rceil \leq NFHH_j x_{jj}$$
$$\forall j \in I$$

Available flights constraints

$$t_{imk} >= 0 \quad \forall i, m \in I, \forall k \in K$$

Linear Mathematical Transformation

Non-linear model: hard to solve the problem in CPLEX

Linear transformation:

$$u_{imj} = x_{ij}\,(1 - x_{mj}) \quad i, j, m \in I.$$

$$O_{imjk} = t_{imk} x_{ij}\,(1 - x_{mj}) \quad i, j, m \in I, k \in K.$$

$$N_{imjk} = t_{imk} x_{ij} \quad i, j, m \in I, k \in K.$$

Modified Objective Function

$$
\begin{aligned}
profit &= revenue - cost \\
&= \sum_{i \in I} \sum_{m \in I} \sum_{k \in K} r_{imk} t_{imk} \\
&- \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{j \in I} \sum_{k \in K} C_{2k} \left(dist_{j0} O_{imjk} + dist_{0j} O_{mijk} \right) \\
&- \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{j \in I} \sum_{k \in K} C_{1k} \left(dist_{ij} N_{imjk} Q2 + dist_{ji} N_{mijk} \right) - \sum_{j \in I} Fixedcost_j x_{jj}
\end{aligned}
$$

New Constraints

$$u_{imj} \geq x_{ij} - x_{mj} \quad i,j,m \in I$$

$$u_{imj} \geq 0 \quad i,j,m \in I$$

$$O_{imjk} \geq t_{imk} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K$$

$$O_{mijk} \geq t_{mik} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K$$

$$N_{imjk} \geq t_{imk} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K$$

$$N_{mijk} \geq t_{mik} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K$$

$$O_{imjk} >= 0 \quad i,j,m \in I, k \in K$$

$$N_{imjk} >= 0 \quad i,j,m \in I, k \in K$$

Stochastic problem

Purpose:

1. maximize the revenue of selling the tickets to different classes and itineraries.
2. minimise the total transportation costs.
3. minimise the total installation costs.

There are scenarios in the stochastic problem.

Input Variables

N	The number of nodes in the hubs-spoke network not considering the central hub.
P	The number of hubs.
K	The number of costumer classes.
$dist_{im}$	Distance from node i to node m.
C_{1k}	unit transfer cost per distance for a customer from class k in a flight between spoke node and hub which is defined as leg type 1.
C_{2k}	unit transfer cost per distance for a customer from class k in a flight between
$nfNH_{ij}$	Number of flights available for itinerary between node i and hub j (each flight travel from node i to j and in the other direction.
$nfHH_j$	Number of flights available for itinerary between hub j and the central hub (each flight travel in both direction).
Q_1	Number of seats available in a flight at leg type 1.
Q_2	Number of seats available in a flight at leg type 2.
$Fixedcost_j$	Fixed cost for establishing non-central hub j.

New Input Variables

S	The number of scenarios.
d^s_{imk}	Traffic demand between origin i and destination m for customer class k under scenario s.
p^s_{imk}	Probability of occurrence of demand d^s_{imk} between origin i and destination m for customer class k under scenario s.

Decision Variables

$t^s_{i,m,k}$ - Int variable. The number of ticket from i to destination j for class k, under scenario s.

$x_{i,j}$ - Binary variable. If there is connection from non-hub i to hub j, the value is equal to 1. If node j is selected as a hub, $x_{j,j}$is equal to 1. The value is 0 in other cases.

$Z_{i,m,k}$ - Int variable. Protection level of the ticket for each class from the origin i to the destination m

FNH_i - Int variable. The number of flight needed to take of from non-hub node i.

FHN_i - Int variable. The number of flight needed to arrive to non-hub node i.

GNH_i - Int variable. The number of flight needed to take of from hub i.

GHN_i - Int variable. The number of flight needed to arrive to hub i.

Objective Function

$$
\begin{aligned}
profit &= revenue - cost \\
&= \sum_{i\in I}\sum_{m\in I}\sum_{k\in K}\sum_{s\in S} t^s_{imk} r_{imk} p^s_{imk} - \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K}\sum_{s\in S} C_{1k}\cdot\left(p^s_{imk} dist_{ij} t^s_{imk} + p^s_{imk} dist_{ji} t^s_{mik}\right) x_{ij} \\
&- \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K}\sum_{s\in S} C_{2k}\cdot\left(p^s_{imk} dist_{j0} t^s_{imk} + p^s_{imk} dist_{0j} t^s_{mik}\right)\left[x_{ij}\left(1 - x_{mj}\right)\right] - \sum_{j\in I} Fixedcost_j x_{jj}
\end{aligned}
$$

Constraints

$$\sum_{j \in I} x_{ij} \leq 1 \quad \forall i \in I$$

$$\sum_{j \in I} x_{jj} = P$$

$$x_{ij} \leq x_{jj} \quad \forall i, j \in I$$

$$t^s_{imk} \leq d^s_{imk} \quad \forall i, m \in I, \forall k \in K, \forall s \in S$$

$$t^s_{\text{imk}} \leq Z_{\text{imk}} \quad \forall i, m \in I, \forall k \in K, \forall s \in S$$

$$\lceil \sum_{m \in I} \sum_{k \in K} \frac{Z_{imk}}{Q1} \rceil \leq \sum_{j \in I} NFNH_{ij} x_{ij} + M x_{ii}$$

$$\lceil \sum_{m \in I} \sum_{k \in K} \frac{Z_{mik}}{Q1} \rceil \leq \sum_{j \in I} NFNH_{ij} x_{ij} + M x_{ii}$$
$$\forall i \in I$$

$$\lceil \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{k \in K} \frac{Z_{imk}}{Q2} x_{ij} \left(1 - x_{mj}\right) \rceil \leq NFHH_{j0} x_{jj}$$

$$\lceil \sum_{i \in I} \sum_{m \in I/\{i\}} \sum_{k \in K} \frac{Z_{mik}}{Q2} x_{ij} \left(1 - x_{mj}\right) \rceil \leq NFHH_{j0} x_{jj}$$
$$\forall j \in I$$

$$t^s_{imk} >= 0$$

$$Z_{imk} >= 0$$
$$\forall i, m \in I, \forall k \in K, \forall s \in S$$

Linear Mathematical Transformation

Non-linear model: hard to solve the problem in CPLEX

Linear transformation:

$$U_{imj} = x_{ij}\left(1 - x_{mj}\right) \quad i, j, m \in I.$$

$$O^{s}_{imjk} = t^{s}_{imk} x_{ij}\left(1 - x_{mj}\right) i, j, m \in I, k \in K, s \in S.$$

$$G_{imjk} = Z_{imk} x_{ij}\left(1 - x_{mj}\right) \quad i, j, m \in I, k \in K.$$

$$N^{s}_{imjk} = t^{s}_{imk} x_{ij} \quad i, j, m \in I, k \in K, s \in S.$$

$$E_{imjk} = Z_{imk} x_{ij} \quad i, j, m \in I, k \in K.$$

Modified Objective Function

$$
\begin{aligned}
profit &= revenue - cost \\
&= \sum_{i\in I}\sum_{m\in I}\sum_{k\in K}\sum_{s\in S} p^s_{imk} r_{imk} t^s_{imk} \\
&- \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K}\sum_{s\in S} C_{2k}\left(p^s_{imk} dist_{j0} O^s_{imjk} + p^s_{mik} dist_{0j} O^s_{mijk}\right) \\
&- \sum_{i\in I}\sum_{m\in I/\{i\}}\sum_{j\in I}\sum_{k\in K}\sum_{s\in S} C_{1k}\left(p^s_{imk} dist_{ij} N^s_{imjk} + p^s_{mik} dist_{ji} N^s_{mijk}\right) - \sum_{j\in I} Fixedcost_j x_{jj}
\end{aligned}
$$

New Constraints

$$u_{imj} \geq x_{ij} - x_{mj} \quad i,j,m \in I$$

$$u_{imj} \geq 0 \quad i,j,m \in I$$

$$O^s_{imjk} \geq t^s_{imk} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K, s \in S$$

$$O^s_{mijk} \geq t^s_{mik} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K, s \in S$$

$$G_{imjk} \geq Z_{imk} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K$$

$$G_{mijk} \geq Z_{mik} - M\,(1 - u_{imj}) \quad i,j,m \in I, k \in K$$

$$N^s_{imjk} \geq t^s_{imk} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K, s \in S$$

$$N^s_{mijk} \geq t^s_{mik} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K, s \in S$$

$$E_{imjk} \geq Z_{imk} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K$$

$$E_{mijk} \geq Z_{mik} - M\,(1 - x_{ij}) \quad i,j,m \in I, k \in K$$

$$O^s_{imjk} >= 0 \quad i,j,m \in I, k \in K, s \in S$$

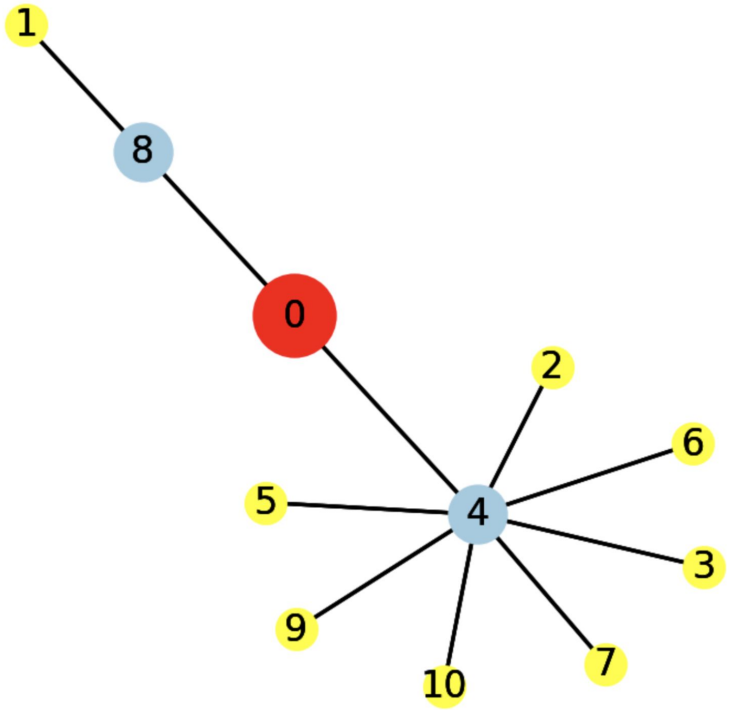
$$N^s_{imjk} >= 0 \quad i,j,m \in I, k \in K, s \in S$$

$$E_{imjk} >= 0 \quad i,j,m \in I, k \in K \quad i,j,m \in I, k \in K$$

$$G_{imjk} >= 0 \quad i,j,m \in I, k \in K$$

Results Presentation

Graphical presentation solved by deterministic model



Best profit: 1884002.59

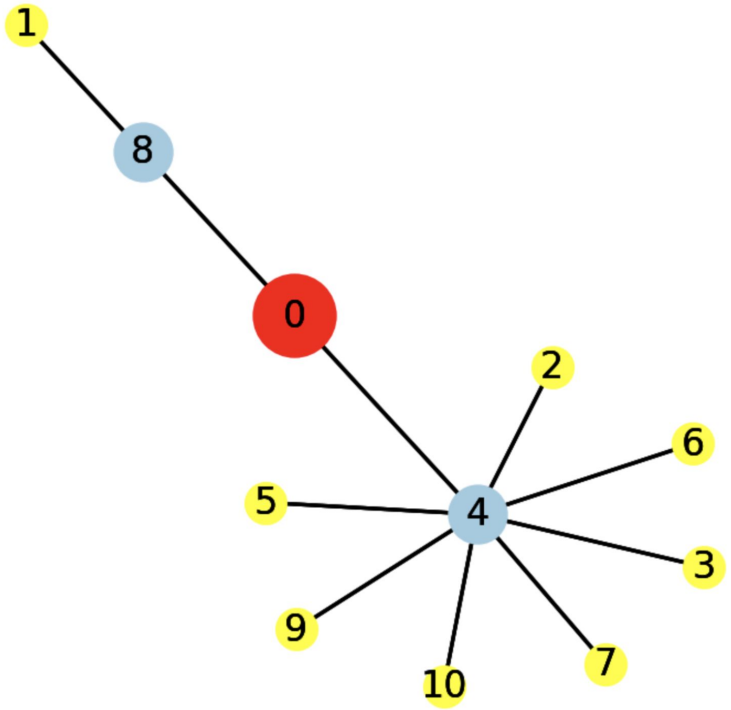
Results Presentation

The result table solved by deterministic model(first 5 rows)

Dest.	pair	class	Sold
1	2	1	10
1	3	1	17
1	4	1	11
1	4	2	47
1	5	1	8

Results Presentation

Graphical presentation solved by stochastic model



Best profit: 1879071.60

Results Presentation

The result table solved by stochastic model(first 5 rows)

Dest.	pair	class	sold1	sold2	prot.
1	2	1	10	10	10
1	3	1	17	11	17
1	4	1	11	15	15
1	4	2	47	47	47
1	5	1	8	18	18

Results Discussion

Case 1: the revenue of tickets for Business class between spoke node 3 and spoke node 4 is drastically increased.

1. The number of ticket sold between node 3 and 4 will be increased. Then the total profit will be increased based on the objective function. (demand is not satisfied)
2. The number of sold tickets will not change, but the total profit will still increase. (demand is satisfied)

Results Discussion

Case 2: the cost discount factor is increased (higher discount)

The node will be distributed more balanced to the two hubs, then the total profit will also increase because of making full use of the number of flights available from hubs to spoke node.

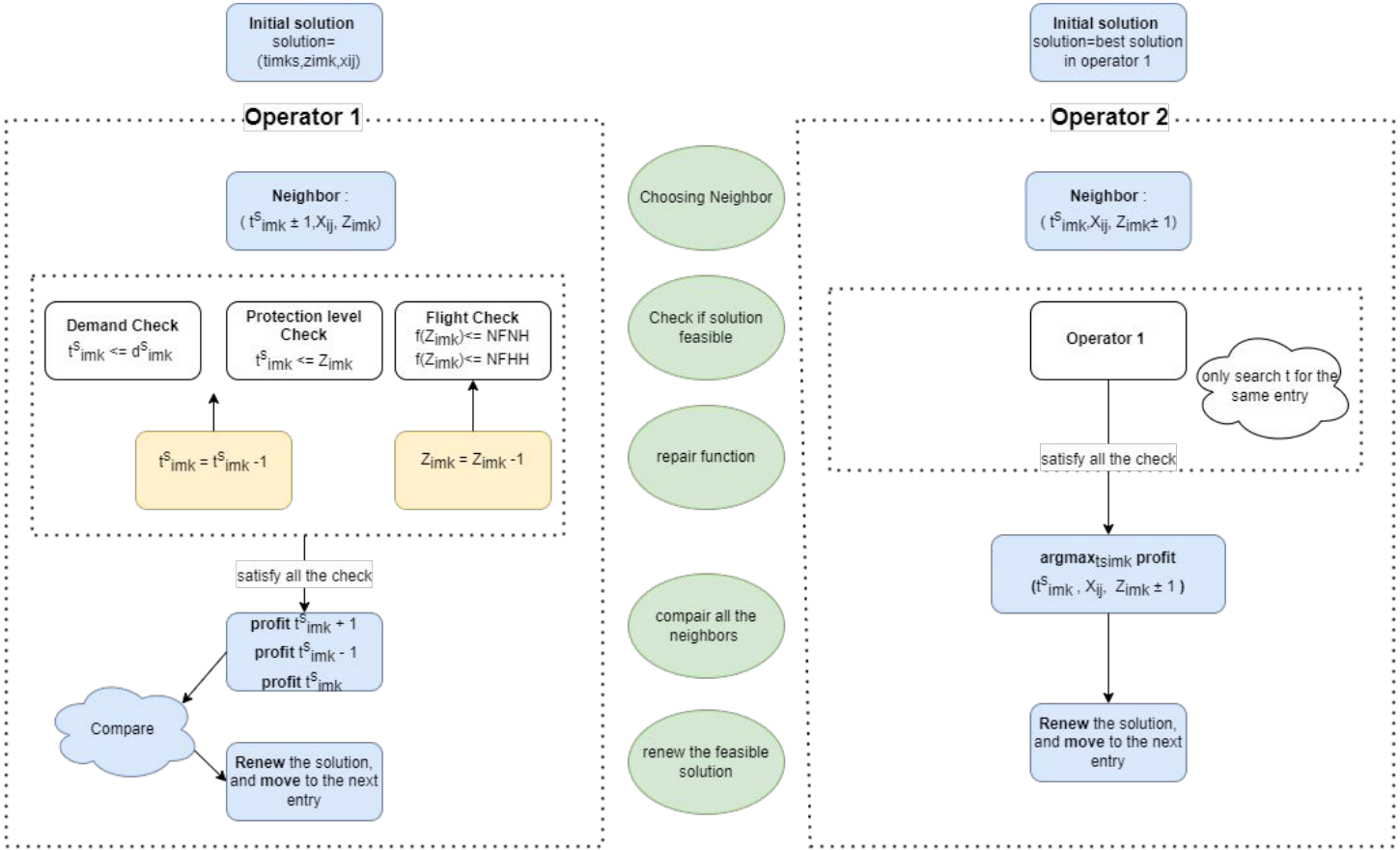
Results Discussion

Case 3: the number of available seats in type 2 legs is increased:

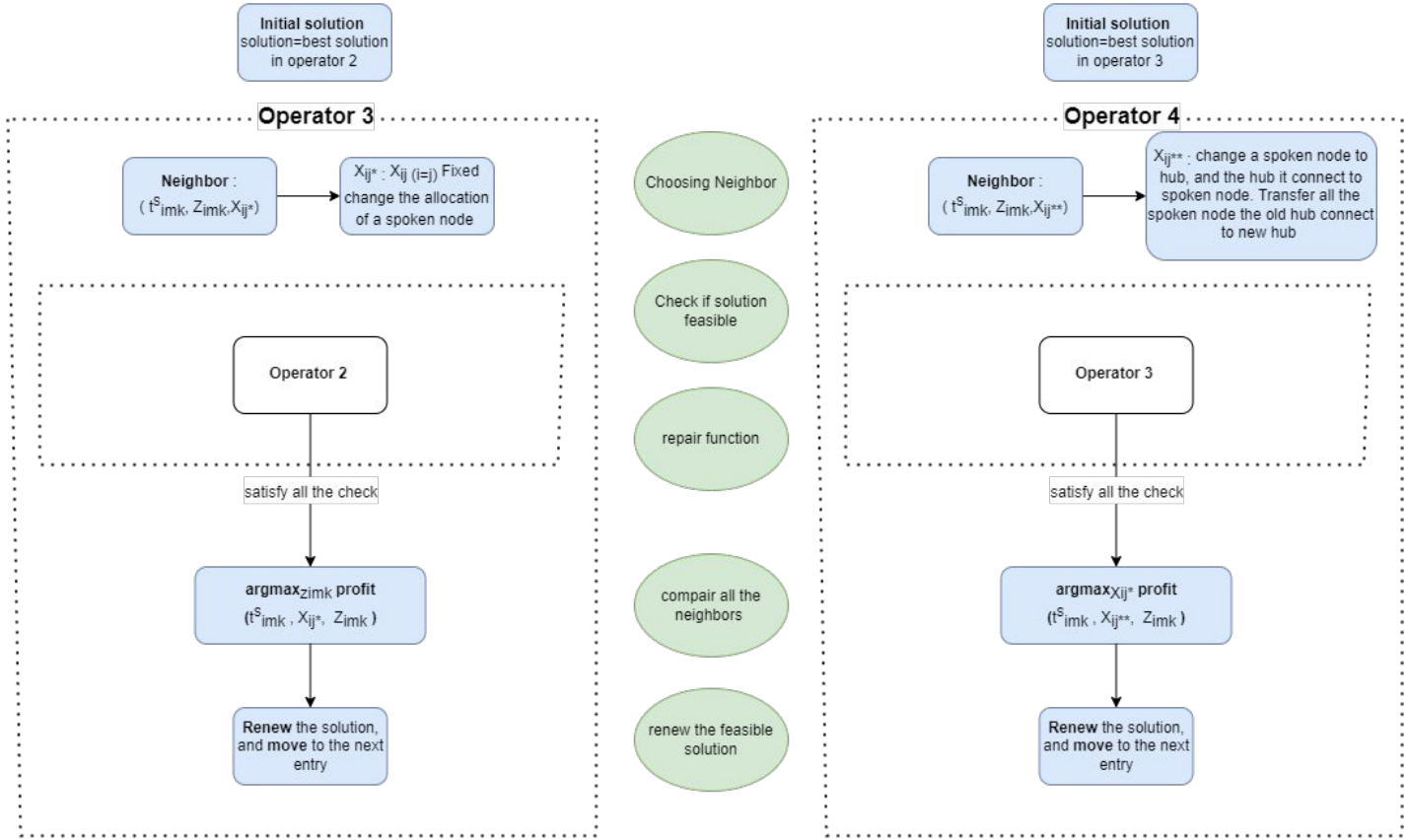
The total profit will increase at the first stage, because there are more passengers who will take the flight crossing the center hub.Then the nodes connected to 2 hubs are more balanced.

If we increase the Q2 until all the demand are satisfied, the total profit will stop increasing.

Meta-heuristic Analysis



Meta-heuristic Analysis



Project 2

Modelling task

Chen Yifeng, Qing Jun

Paper Review

For MNL (multinomial logit) method:

• Advantages:

– Low computational cost;
– Compared to the SVM method, it can obtain higher accuracy;
– Compared to the SVM and ANN method, it can obtain lower standard deviation.

• Disadvantage :

– Compared the RF, BOOST, NB, KNN, BAG, CART, it obtains lower accuracy;
– Compared the RF, BOOST, NB, KNN, BAG, CART, it obtains higher standard deviation.

Paper Review

For machine learning approaches:

• Advantages:

– For RF, BOOST, NB, KNN, BAG, CART method, they can acquire higher accuracy for prediction
– For RF, BAG, CART and KNN methods, the accuracy can be improved if the sample size is increasing
– For RF, BOOST, NB, KNN, BAG, CART method, they can have smaller standard deviation.

• Disadvantage :

– For SVM method, since it is linear classification, the prediction for non-linear classification problem is not good.
– High computational cost.

Paper Review

Improving suggestions:

• Model hyper-parameters selection method: The reviewed paper do not perform the search of hyper parameters inside the models. But for choose the models, they use different train-test data split ratios;

• Validation Scheme: They use K fold validation for comparing with k of 10,20,and 30 separately;

• Sampling strategy in validation: After comparing the models under different split ratios and cross validation. The author evaluates the nine models accuracy under different sampling size;

• Performance metrics: The author uses accuracy, precision, recall, and F1 score to evaluate the performance of different models;

• Data imbalance : The data imbalance is left to be processed in the reviewed paper

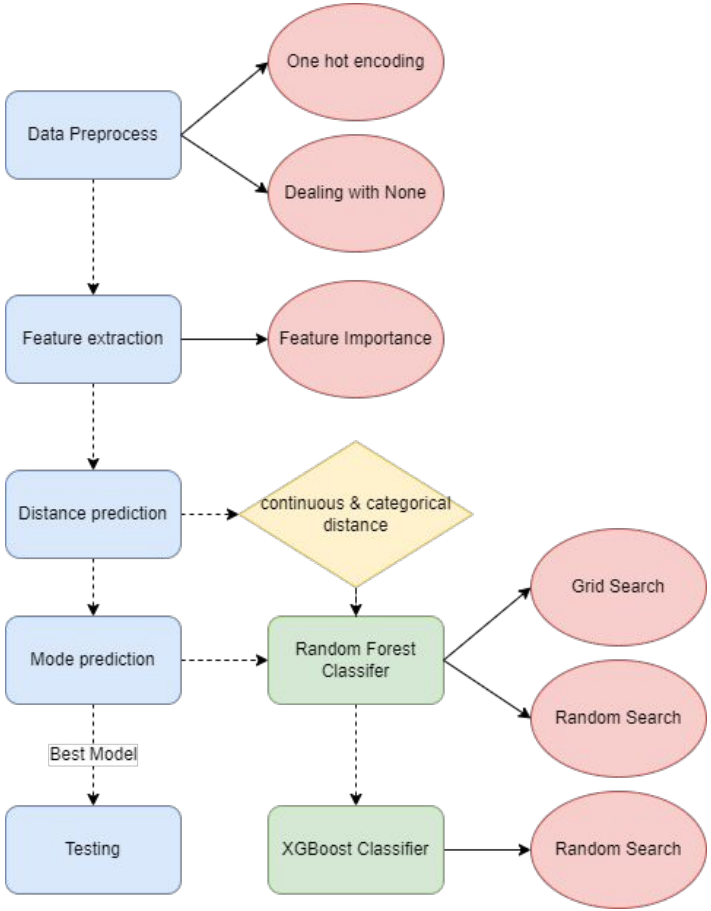
Purpose

Predict trip distance

mode choice prediction

Column type	Column name
Time	travel_date
Categorical	travel_date_dow, o_purpose_category, d_purpose_category, o_location_type, d_location_type o_congestion, d_congestion,age, employment, student,license, planning_apps, disability industry, gender, education, survey_language, res_type, rent_own, income_aggregate
Numerical	num_non_hh_travelers, num_hh_travelers, num_travelers, num_bicycles, num_vehicles, num_people, num_adults, num_kids, num_workers, num_students
Meaningless	hh_id, person_id, trip_id
Intermediate target	trip_distance(numerical)
Target	mode(categorical)

Flow chart


One hot encoding
Data Preprocess
Dealing with None
Feature extraction
Feature Importance
Distance prediction
continuous & categorical distance
Grid Search
Mode prediction
Random Forest Classifer
Random Search
Best Model
Testing
XGBoost Classifier
Random Search

Preprocess

1 Changing the type of ‘time’

```
date_time = pd.to_datetime(data['travel_date'], format='%d/%m/%Y').dt.dayofweek # change week time
date_time[date_time < 5] = 0
date_time[date_time >= 5] = 1
data['travel_date'] = date_time
```

Preprocess

2 Dealing with categorical column

one-hot encoding

```
def Get_D(name,data):
    res = pd.get_dummies(data[name]).rename(columns=lambda x:name+'_' +str(x))
    data = data.join(res)
    data.drop(columns=[name], inplace=True)
    return data
```

Preprocess

3 Dealing with the empty values - proposed ways

3.1 Replace the column with mean

3.2 Replace the column with the distribution of other data

3.3 Create a new feature for the empty value

```
length1 = len(data[data['rent_own'] != -9998]['rent_own'])/len(data[data['rent_own'] == -9998]['rent_own'])
len_none = len(data[data['rent_own'] == -9998]['rent_own'])
data[data['rent_own'] != -9998]['rent_own'].value_counts()/length1
list_rent_type_num = [2 for i in range(2459)] + [1 for i in range(1571)] + [3 for i in range(11)]
list_index_none_rent_type = list(data[data['rent_own'] == -9998].id)
for index, i in enumerate(list_index_none_rent_type):
    data['rent_own'][i] = list_rent_type_num[index]
```

Preprocess

4 Dealing with the imbalanced features

There are several features with fewer samples, for example, in the gender column, there is a non-binary feature contributes only 0.4% samples among all the samples, which could lead to overfitting problem.

```
#delete the imbalanced data
search_columns_feature1=search_columns_feature.copy()
for i in search_columns_feature1:
    if data[i].mean()<0.05:
        search_columns_feature=search_columns_feature.drop(i)
```

Preprocess

5 Standardization

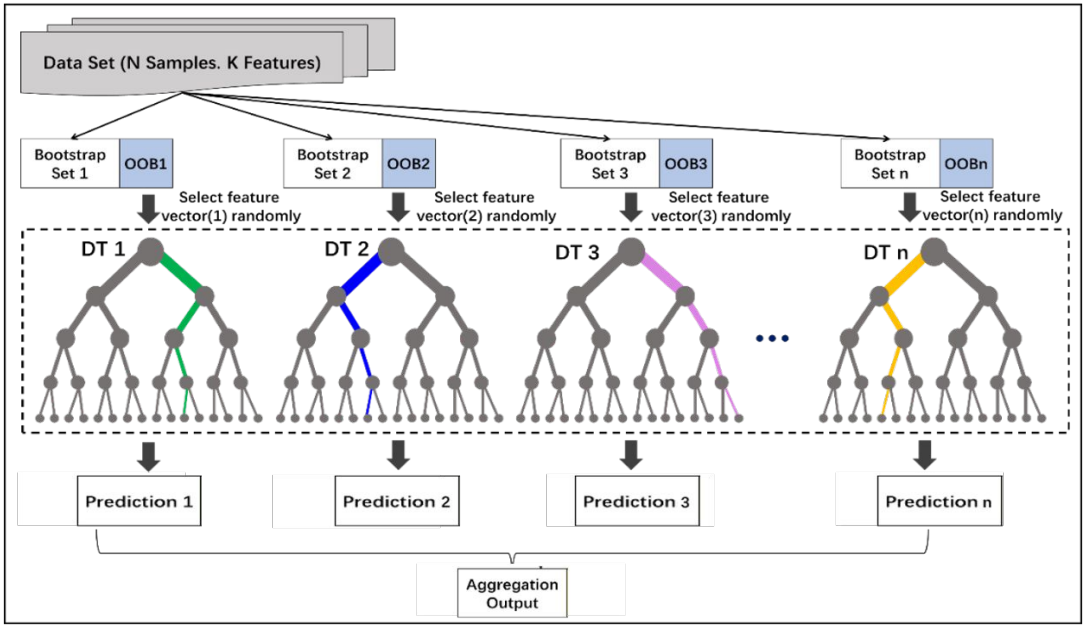
Before fitting the features to the model, we must standerlize the input to ensure the scale of the value do not impact the weight.

$$x_{scaled} = \frac{x - \mathbf{x}_{\text{mean}}}{std}$$

```
sc = StandardScaler()
sc.fit(X_train)
X_train_std = sc.transform(X_train)
X_test_std = sc.transform(X_test)
```

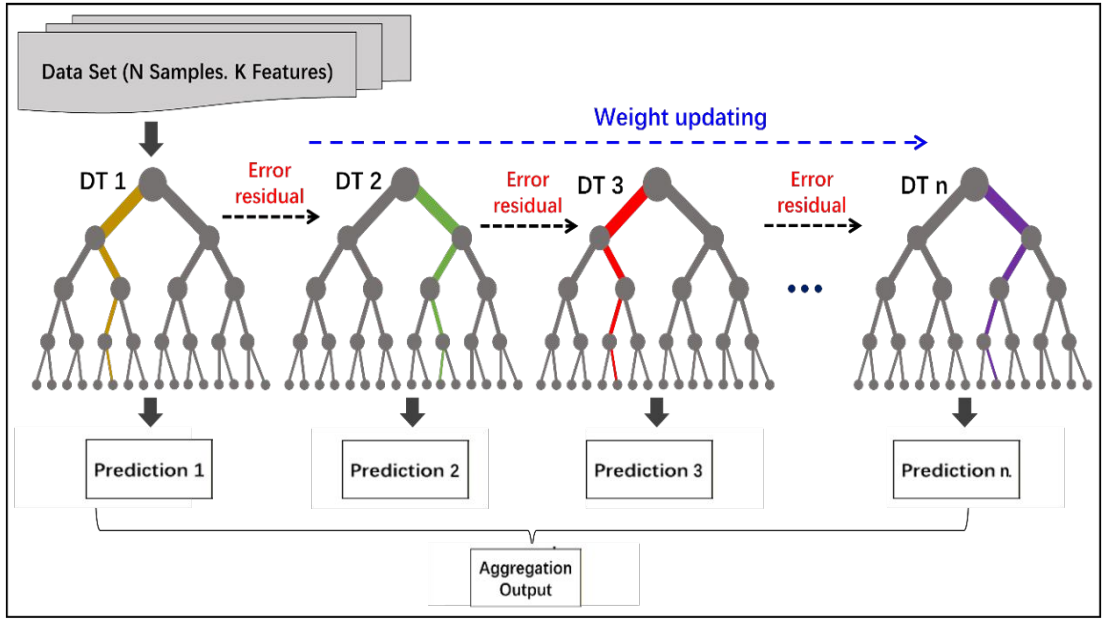
Model choice

Random Forest



Model choice

XGBoost



Hyperparameter search

1 For hyper parameter in model - Cross Validation

Grid Search, Random Search

2 Distance type

Continuous, Categorical

3 Number of features

PCA importance for feature
Feature importance from model

Hyperparameter search

1 For hyper parameter in model - Cross Validation

Grid Search, Random Search

```
clf = RandomForestRegressor(random_state=0)
grid = GridSearchCV(clf, rf_params, cv=5, scoring='neg_log_loss')
grid.fit(X_train_std,y_train['trip_distance'])

clf = RandomForestRegressor()
Random = RandomizedSearchCV(clf, param_distributions=rf_params,cv=5,n_iter=20)
Random.fit(X_train_std,y_train['trip_distance'])
```

Hyperparameter search

2 Distance type

Continuous, Categorical

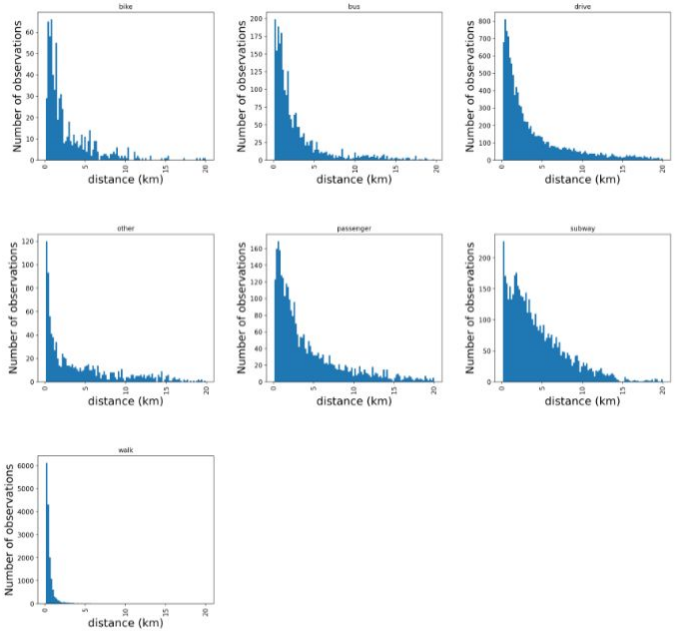
```python
def transfrom_distance(x):
    """change the format of distance

    Args:
        x (float): the value of distance

    Returns:
        int: the category value of distance
    """
    if x < 3:
        return 0

    elif x<4:
        return 1

    else: return 2
```



Hyperparameter search

3 Number of features

PCA importance for feature

```
# Perform PCA
pca = PCA()
pca.fit(X_train_std)

ranks = pd.DataFrame(pca.components_, columns=list(X_train.columns)).loc[len(search_columns)-1]
ranks_full = abs(ranks).sort_values(ascending=False)
ranks_full.to_csv('ranks.csv')
search_columns_pca = abs(ranks).sort_values(ascending=False)[:num_features].index
```

Hyperparameter search

o_purpose_category_-9998	0.299370
industry_11	0.287322
industry_1	0.265045
industry_16	0.251437
d_purpose_category_10	0.214095
	...
d_location_type_4	0.003293
age_4	0.001002
d_location_type_2	0.000750
res_type_-9998	0.000678
num_workers	0.000007

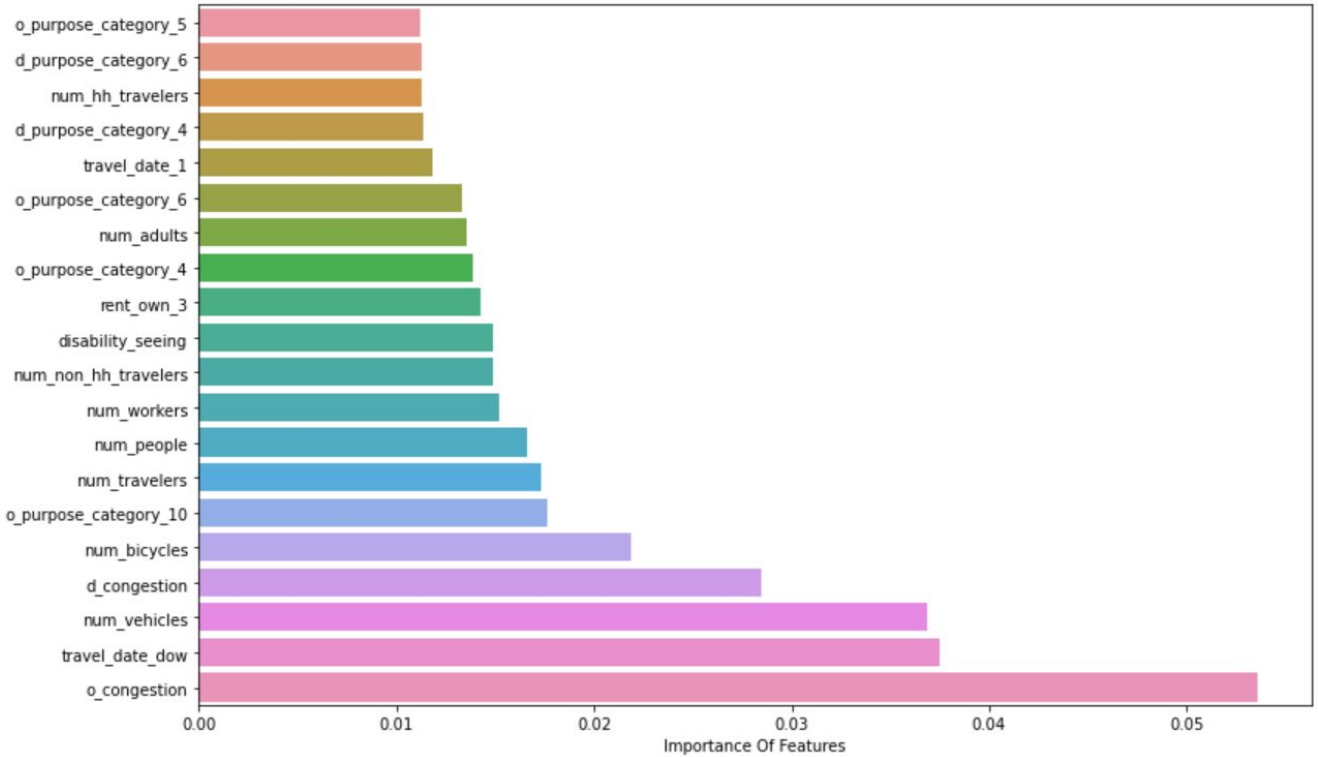
Hyperparameter search

3 Number of features

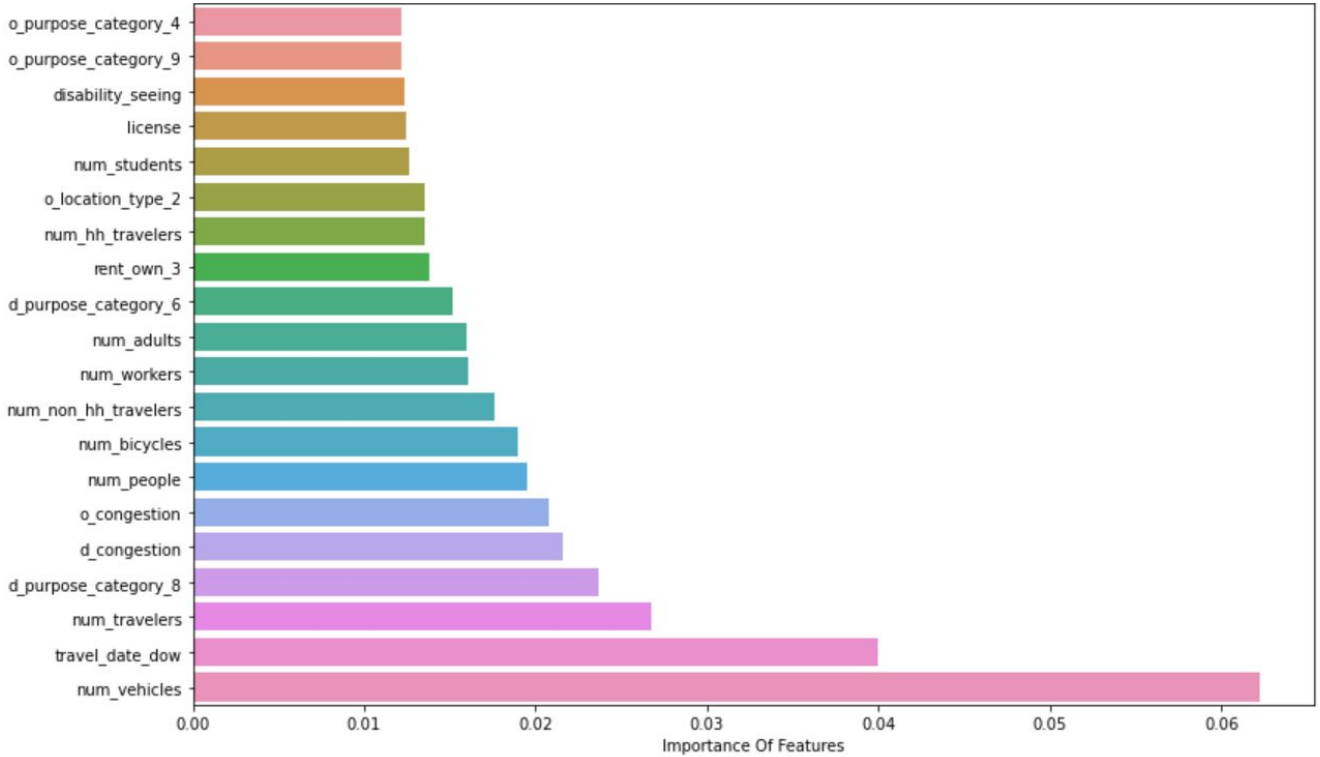
Feature importance from model

```
rf = RandomForestRegressor()
rf.fit(X,y)
data_tuples = list(zip(search_columns,rf.feature_importances_))
df=pd.DataFrame(data_tuples, columns=['name','value'])
dataindex=list(df.sort_values('value',ascending=False)[0:num_features]['name'].values)
dataindex=pd.Index(dataindex)
# dataindex.to_csv('features.csv')
```

Hyperparameter search



Hyperparameter search



Model comparison and selection

1 Predict trip distance

evaluation:

continuous distance:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N}}$$

categorical distance:

$$loss = -\sum_{n=1}^{N} \ln P(i_n|x_n)$$

Model comparison and selection

2 mode choice prediction

model:

random forest grid search

random forest random search

XGBoost with random search

Regression model	Number of features	60	80	100	all
Distance Prediction (loss)	Grid search Random Forest	3.3047	3.2973	3.2979	3.2968
	Random search Random Forest	3.3055	3.2973	3.2980	3.2967
Mode Prediction (cross entropy)	Grid search Random Forest	1.1020	1.1113	1.1144	1.1133
Based on the distance predicted	Random search Random Forest	1.1032	1.1088	1.1123	1.1133
by all parameters	Random search XGBoost	0.7704	0.7462	0.7373	0.7405

Classification model	Number of features	60	80	100	all
Distance Prediction (cross entropy)	Grid search Random Forest	0.6492	0.6492	0.6504	0.6540
	Random search Random Forest	0.6467	0.6509	0.6477	0.6517
Mode Prediction (cross entropy)	Grid search Random Forest	1.2550	1.2000	1.2090	1.3570
Based on the distance predicted	Random search Random Forest	1.257 0	1.1960	1.2100	1.1350
by best 60 parameters	Random search XGBoost	0.7712	0.7653	0.7452	0.7388

Model comparison and selection