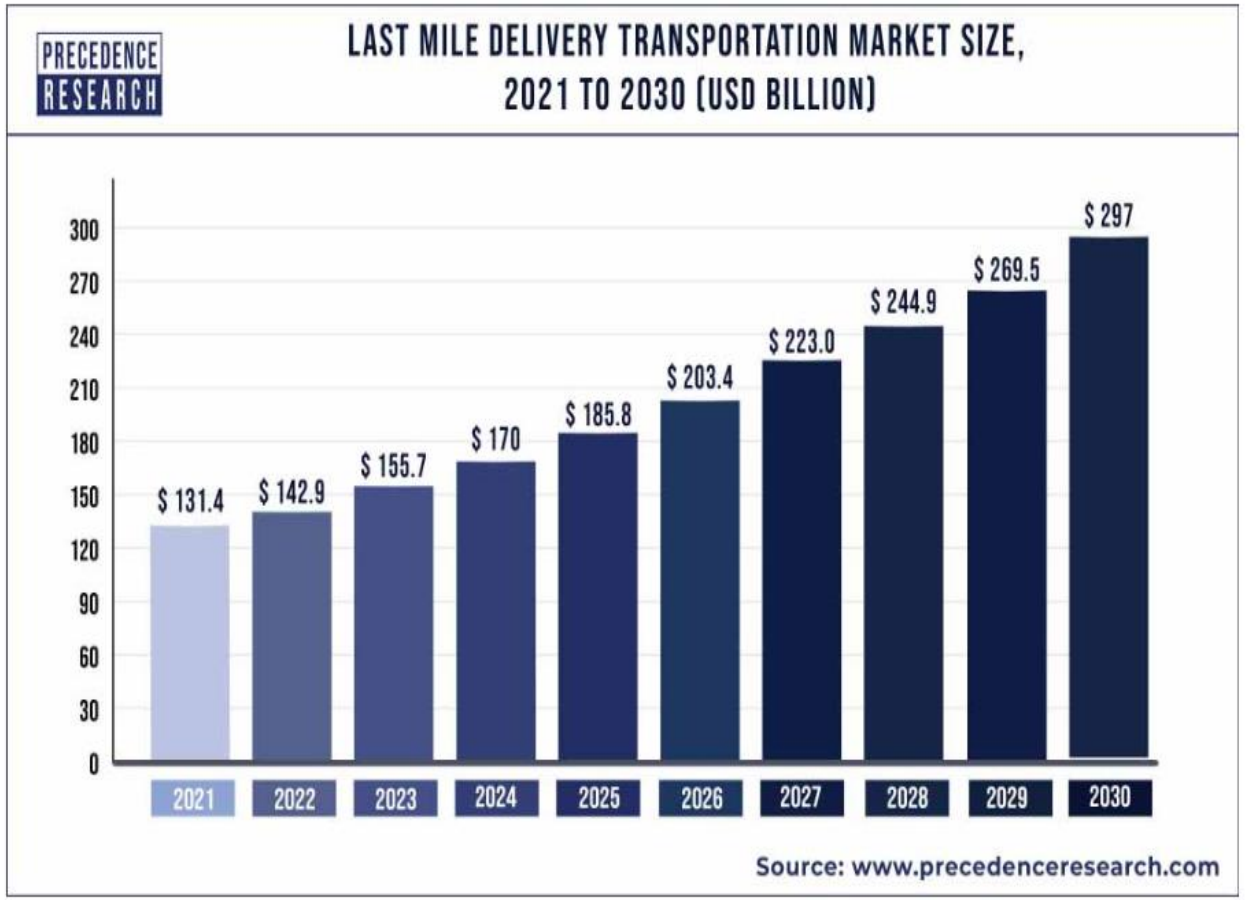
Rise in Last-Mile Delivery

Revolutionary era of rapid delivery to **our doorstep**

Online sales surged due to the COVID-19 pandemic, leading to a significant rise in e-commerce companies

E-commerce and retail delivery services heavily rely on the **Last-Mile delivery system**

The last-mile delivery industry is projected to grow at a **9.6% CAGR** (Compound Annual Growth Rate) from 2022 to 2030



Last-mile Delivery Driver

- **Frontline workers** in last-mile logistics
- Logistics companies provide their drivers with recommended optimized routes to complete their deliveries
- Frequently **deviations** from these suggested routes are observed **in actual delivery cases**
- Despite numerous research and improvements in last-mile delivery solutions employing various route optimization algo rithms,the practical application of these models is still **challengin g due to the influence of many real-world factors on the ground**

https://reliantcouriers.co.uk/large-items/

Delivery Driver Behaviour

- Dang Bao Le (2022) stated that experienced drivers have **implicit knowledge of local conditions** that is hard to quantify in optimization models, and **these factors all influence** how well a planned optimized route will be **executed on the real ground** [1].
- According to the study of Dai et al. (2016), drivers choose alternative routes for a particular route based on the variables such as **traffic conditions, weather, road maintenance, or even their own mood** [2].
- Srivatsa and Gajanand (2017) did empirical studies and found that **in non-congested settings**, drivers typically adhere to the planned route,while **in congested settings**, route-choice behaviour is more affected by factors such **as recurrent congestion (traffic flow)** and **non- recurrent congestion(sudden incidents)** [4] **.**

Research Objective

Analyse and predict the **driver's behaviour** in case of compliance or deviation with
the suggested optimized routes **in the last-mile delivery**

Identify patterns and trends in driver behaviour through **data visualization**

Develop **quantitative metric** to measure on compliance

Employ and formulate **modelling approaches** to predict binary decision on compliance

Quantify and compare the **effects of the relevant factors** that influence drivers' decisions regarding compliance

Actual Last-mile Delivery Dataset

- We used **historical data** from actual driving routes within the last-mile delivery plans of **a large e-commerce parcel delivery company.**
- The company offers effective **last-mile delivery solutions** to its retailer clients, enabling them to provide delivery services to their customers.
- The dataset consists of **daily last-mile delivery records** of this company **for the year 2022.**
- These records pertain to the delivery activities **of their professional drivers** operating from two of their **major warehouses located in Lisbon and Barcelona**.

Raw Dataset

Shipments	DataType	Data Description
shipment_id	Integer	unique shipment identifier
order_id	Integer	unique retailer order id (there can be multiple shipments for the same order_id)
route_id	Integer	Unique id of route where the shipment was planned to be performed
latitude_delivery	Float	latitude of delivery location
longitude_delivery	Float	longitude of delivery location
timewindow_start	Datetime	start of time slot for delivery in format DD/MM/YYYY HH:MM:SS
timewindow_end	Datetime	end of time slot for delivery in format DD/MM/YYYY HH:MM:SS
latitude_attempted_time	Float	estimated driver's latitude at the time of the attempt, if any
longitude_attempted_time	Float	estimated driver longitude at the time of the attempt, if any
company_id	Integer	unique identifier of the retailer who placed the order
Attempted	Bool	1 if attempted, 0 otherwise
Delivered	Bool	1 if successfully delivered, 0 otherwise
Country	String	country name
City	String	city name
driver_id	Integer	unique driver identifier
scheduled_sort_order	Integer	planned route position within the sequence of deliveries represented by ordered numbers(1,2,3,...)
attempted_time	Datetime	time at which delivery was attempted if attempted in format DD/MM/YYYY HH:MM:

Data Cleaning,Filtering and Imputation

- Dropped the rows corresponding to **the missing values** in 'route_id' , 'scheduled_sort_order', or 'attempted_time'
- Box plots to **identify and remove outliers** in latitude and longitude values
- Used KNN(K-Nearest Neighbour) technique **to impute the missing data** for ‘latitude_attempted_time’ and ‘longitude_attempted_time’
- **Removed duplicates** with identical delivery locations and attempted times
- Refined the raw data into a precise format for in-depth analysis resulting in approximately **600,000 delivery orders from around 10,000 distinct routes** for each warehouse

Feature Engineering

- 'tw_month' and 'tw_duration‘ were generated from the 'timewindow_start' and 'timewindow_end' columns
- **Haversine formula (1)** to calculate the distance between the **consecutive stops in the routes**,

$$a = sin^2\left(\frac{\Delta lat}{2}\right) + cos\, lat1 \cdot cos\, lat2 \cdot sin^2\left(\frac{\Delta lon}{2}\right)$$

$$c = 2 \cdot tan^{-1}\left(\frac{\sqrt{(a)}}{\sqrt{1-a}}\right)$$

$$distance = R\,.c \qquad (1)$$

- The **total planned distance** for each route_id was then **determined by adding these consecutive distances**

Defining Compliance Metric

- Defined **a new metric called 'stop_compliance'** which takes binary values (0 and 1) to indicate whether a driver visits **a specific delivery stop in the suggested sequence order in a route or not**
- Created **a new column named 'actual_order'** assigning chronological numbers(1,2,3,…) **to represent the actual delivery sequence** by sorting the 'attempted_time' column in ascending time within each route
- Compared the driver's actual delivery sequence i.e. ‘actual_order’ with the suggested delivery sequence i.e ‘scheduled_sort_order’ **to determine the compliance for each delivery stop in the routes**

Defining Compliance Metric

The stop_compliance variable denoted by C_{ri} **evaluates pairwise stop compliances** in route subsequence using the following formula,

$$C_{ri} = 1 \quad when \{ \left|(i-1)_{r(i-1)} - (i)_{ri}\right| - \left|(j-1)_{r(i-1)} - (j)_{ri}\right| = 0$$

$$or,$$

$$\left|(i-1)_{r(i-1)} - (i)_{ri}\right| - \left|(j+1)_{r(i-1)} - (j)_{ri}\right| = 0$$

$$or,$$

$$\left|(i+1)_{r(i+1)} - (i)_{ri}\right| - \left|(j+1)_{r(i+1)} - (j)_{ri}\right| = 0$$

$$or,$$

$$\left|(i+1)_{r(i+1)} - (i)_{ri}\right| - \left|(j-1)_{r(i+1)} - (j)_{ri}\right| = 0 \}$$

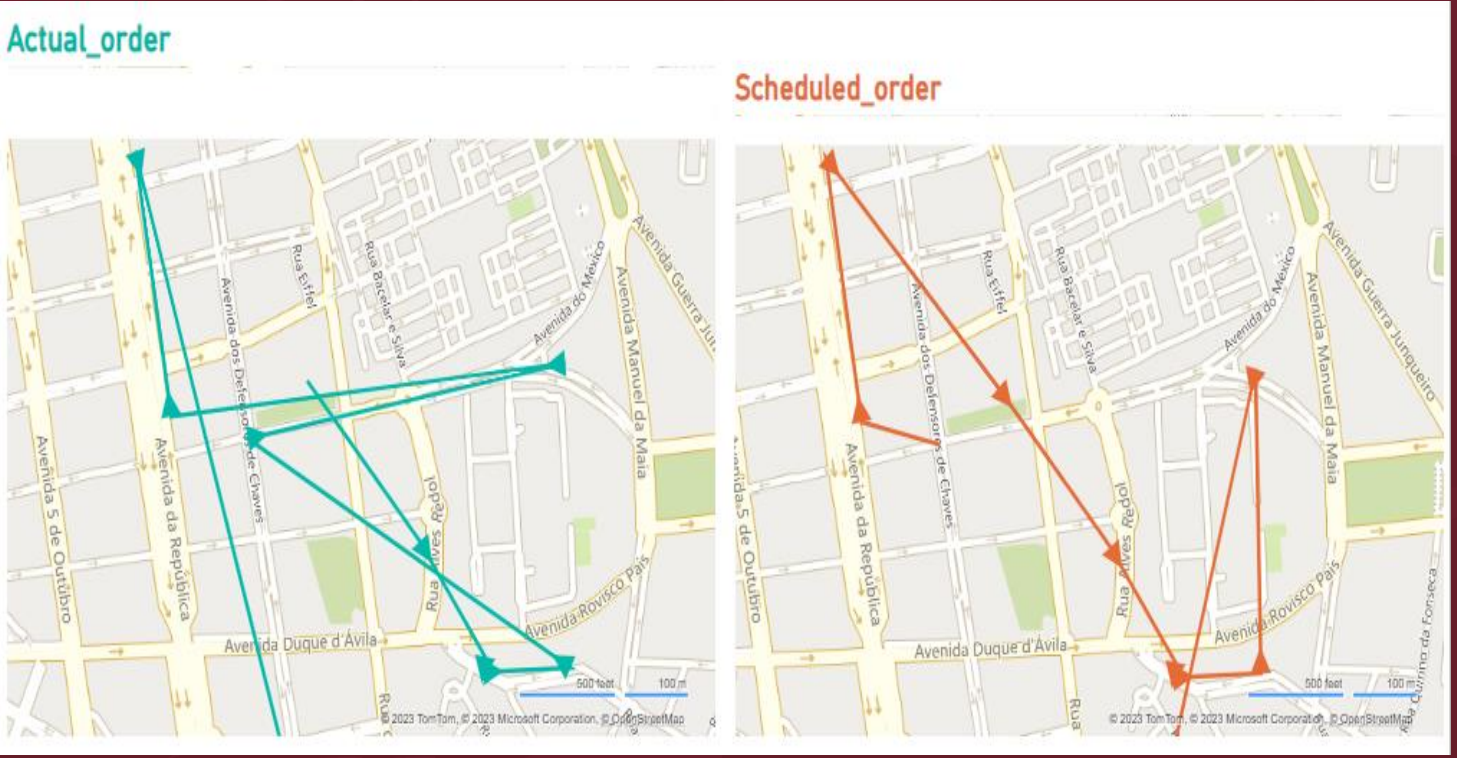
$$C_{ri} = 0, \; Otherwise \qquad (2)$$

The formula (2) assigned a **value of 1** to the corresponding 'stop_compliance' column for **two consecutive pair of stops** if for any route instance r, there are two adjacent suggested stops denoted as j_r and $(j+1)_r$ or $(j-1)_r$ and j_r, such that the driver **follows these adjacent stops** regardless of the direction to deliver products,by comparing with actual order index denoted as i_r and $(i+1)_r$ or i_r and $(i-1)_r$. If not, a **value of 0** is assigned.

Compliance from Forward Direction

The compliance metric evaluates whether a driver adheres to two adjacent scheduled orders, **irrespective of the direction** they are following—be it from **the front or the back**.

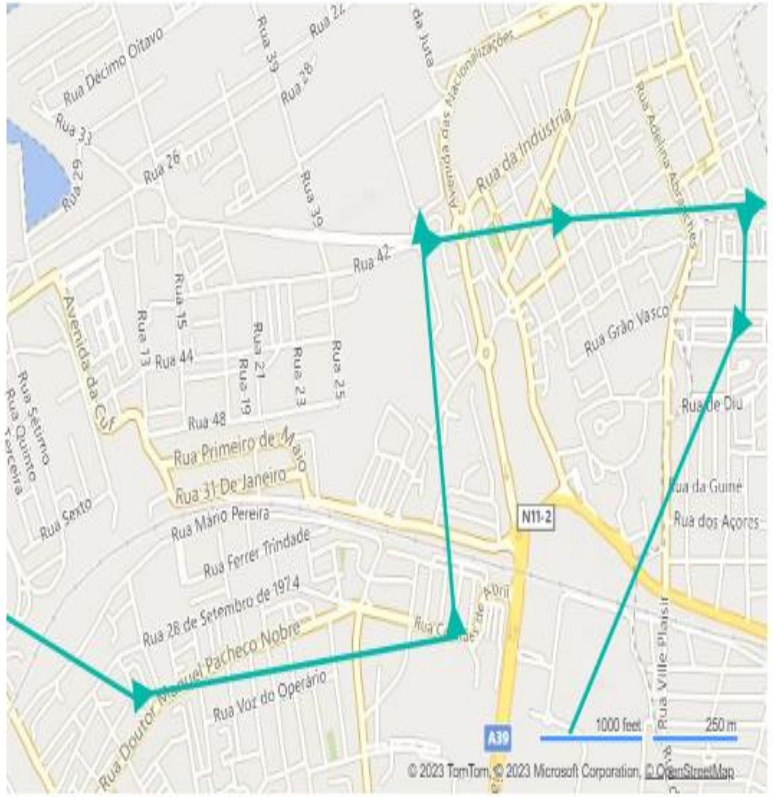
	route_id	order_id	scheduled_sort_order	actual_order	stop_compliance
58827	996545.0	448819	69.0	47	1
58828	996545.0	294344	70.0	48	1
58829	996545.0	477874	71.0	49	1
58830	996545.0	400604	72.0	50	1
58831	996545.0	65009	73.0	51	1
58832	996545.0	595385	48.0	52	0
58833	996545.0	560149	74.0	53	0
58834	996545.0	341143	49.0	54	1
58835	996545.0	583352	50.0	55	1



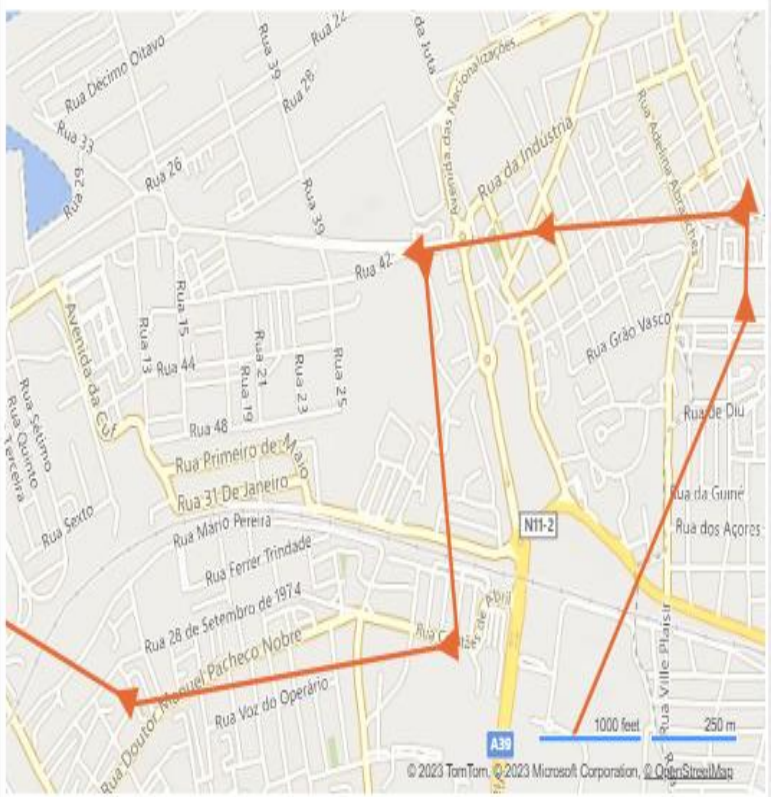
Compliance from Backward Direction

	route_id	order_id	scheduled_sort_order	actual_order	stop_compliance
58877	997607.0	428330	47.0	38	1
58878	997607.0	279218	46.0	39	1
58879	997607.0	611269	45.0	40	1
58880	997607.0	352493	44.0	41	1
58881	997607.0	589681	43.0	42	1
58882	997607.0	503289	42.0	43	1
58883	997607.0	309688	41.0	44	1
58884	997607.0	338438	40.0	45	1
58885	997607.0	612300	39.0	46	1
58886	997607.0	334266	38.0	47	1

Actual_order



Scheduled_order



Additional Features on Demographic, Weather and Urban Morphology

Explanatory Features

Route Level Features

- Time-window duration
- Stop count
- Company count
- Municipality count
- Total planned distance

Municipality Level Features

- Population Density (km^2)
- Number of trunk roads
- Intersection_count
- Average speed limit
- Street length average
- Average daily temperature
- Average daily precipitation

Route Category

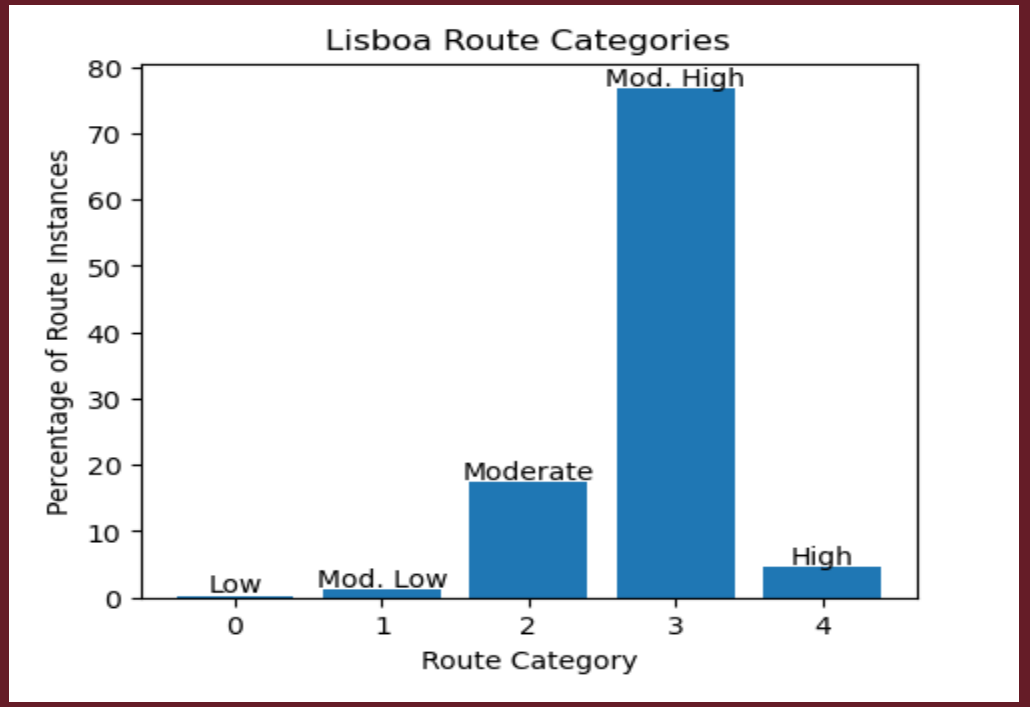
- 'compliance_ratio' represents **the proportion of stops followed by a driver** in a route is given by,

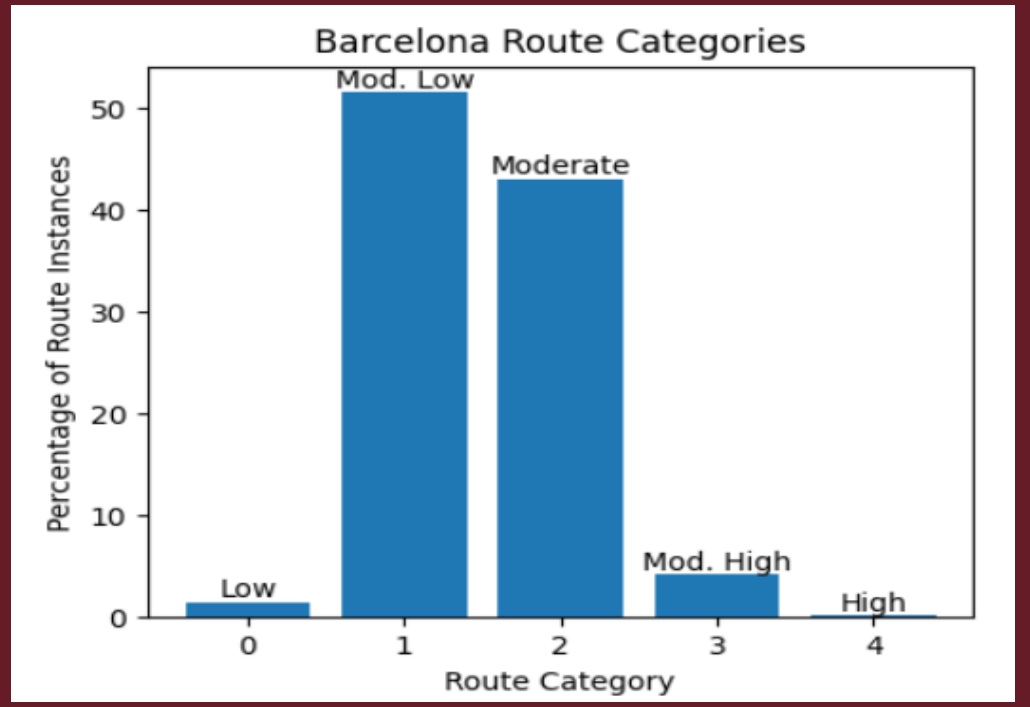
$$compliance_{ratio} = \frac{total\ followed\ delivery\ stops\ in\ a\ route}{total\ delivery\ stops\ in\ a\ route} = \frac{\sum_{i=1}^{N_r} C_{ri}}{N_r}$$

$$compliance_{rate} = compliance_{ratio} * 100$$

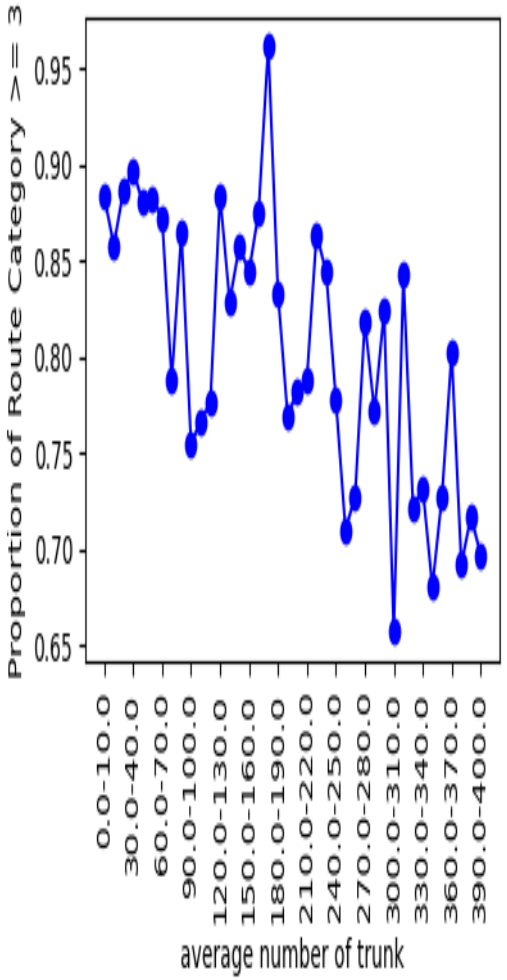
- **Category 4** =$compliance_{rate}$=100% = **High Category Route**
- **Category 3** =$70\% \leq compliance_{rate} < 100\%$ = **Moderately High Category Route**
- **Category 2** =$40\% \leq compliance_{rate} < 70\%$ = **Moderate Category Route**
- **Category 1** = $0\% < compliance_{rate} < 40\%$ = **Moderately Low Category Route**
- **Category 0** = $compliance_{rate}$= 0% = **Low Category Route**

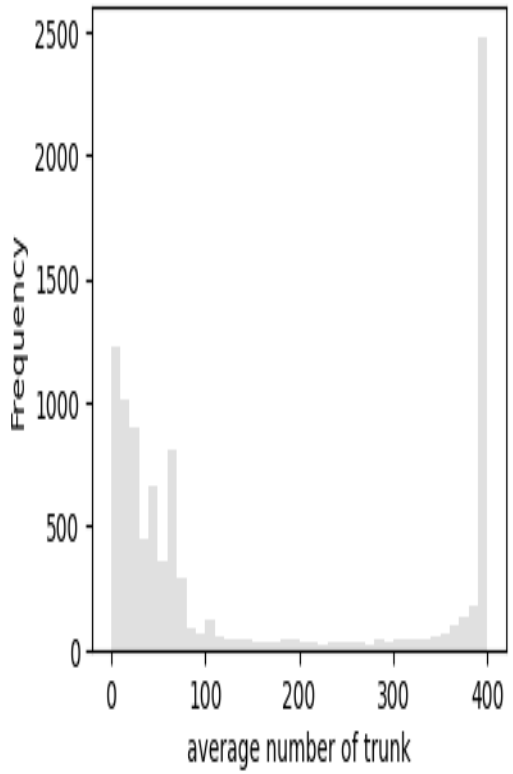
Route Category

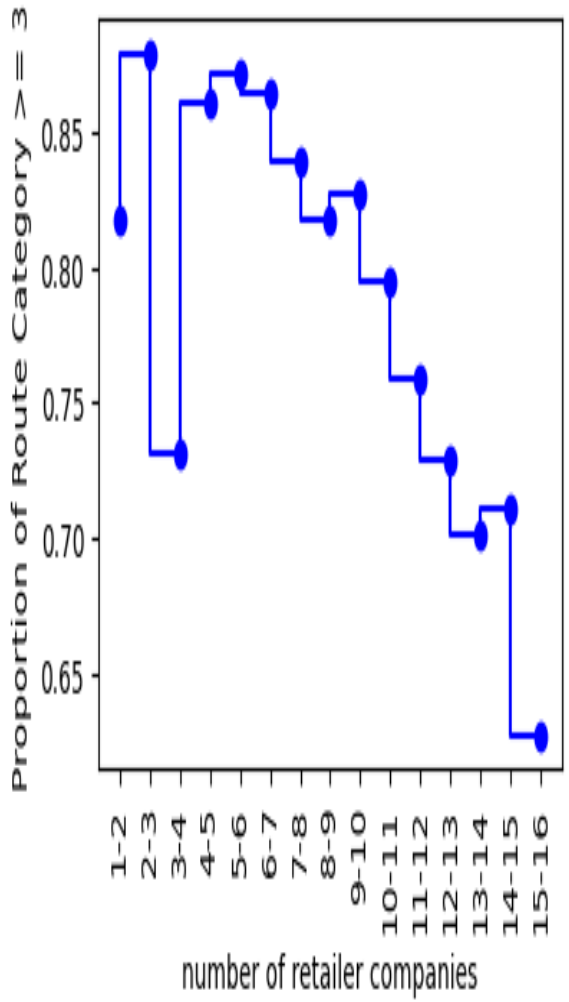


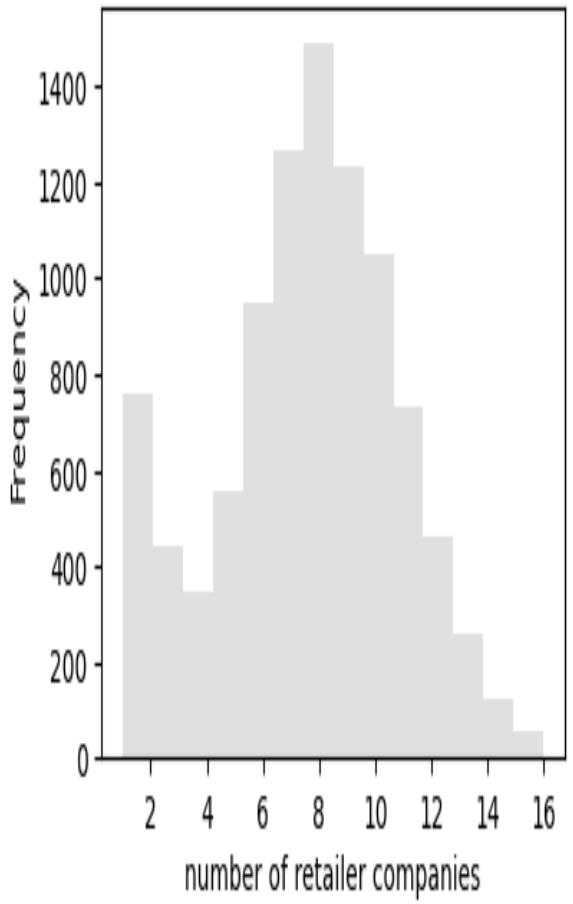


EDA on Lisboa Dataset

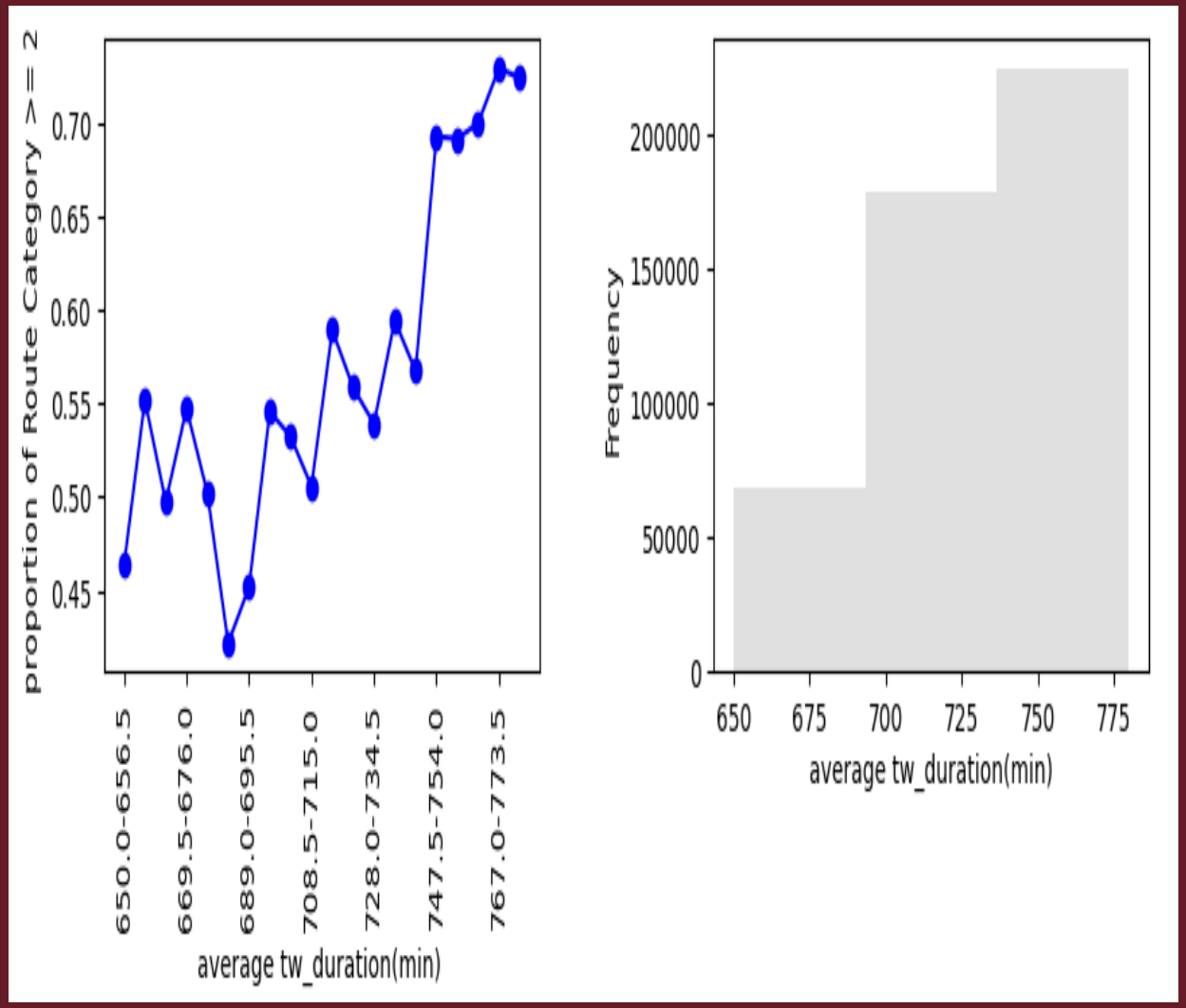


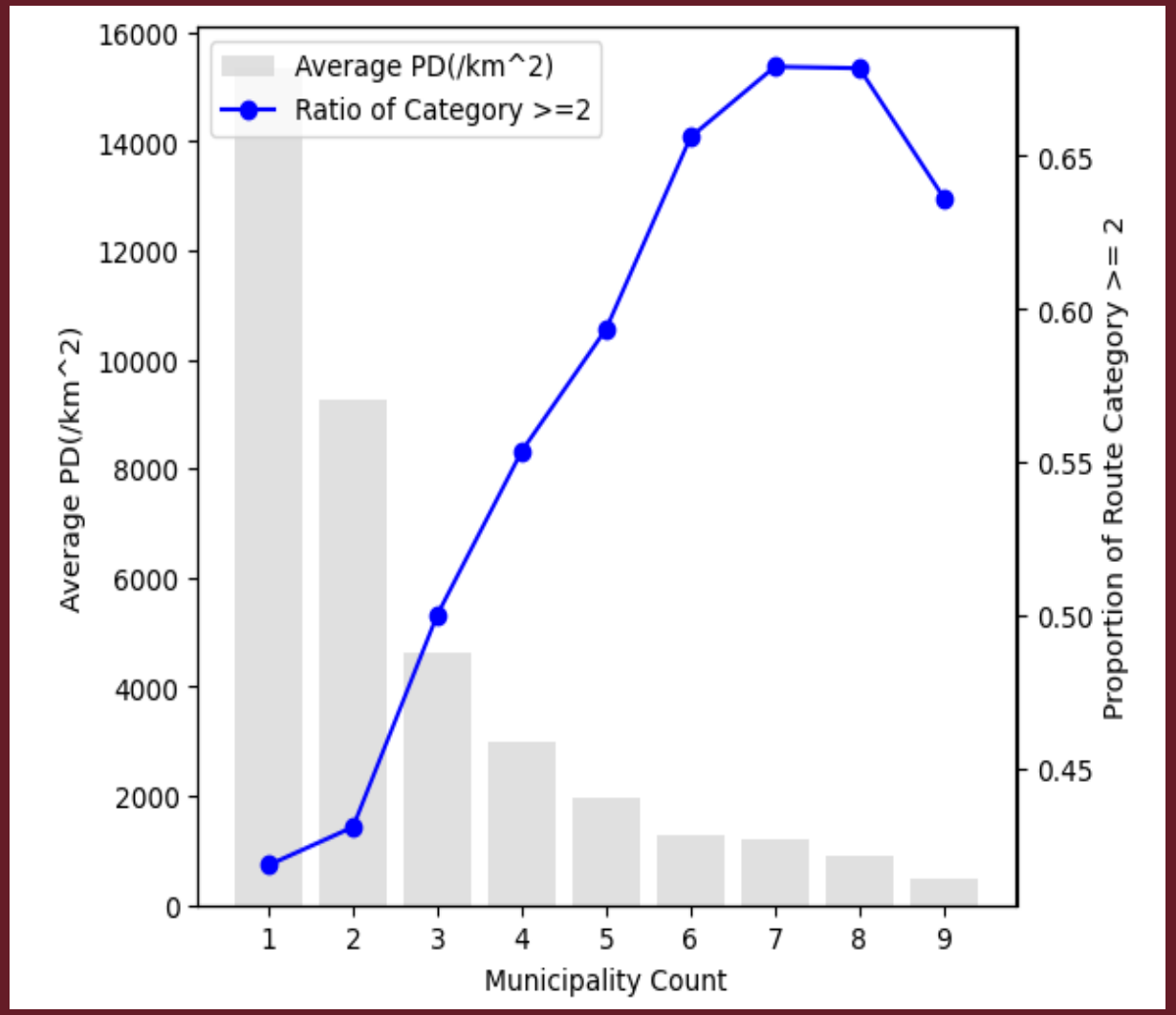




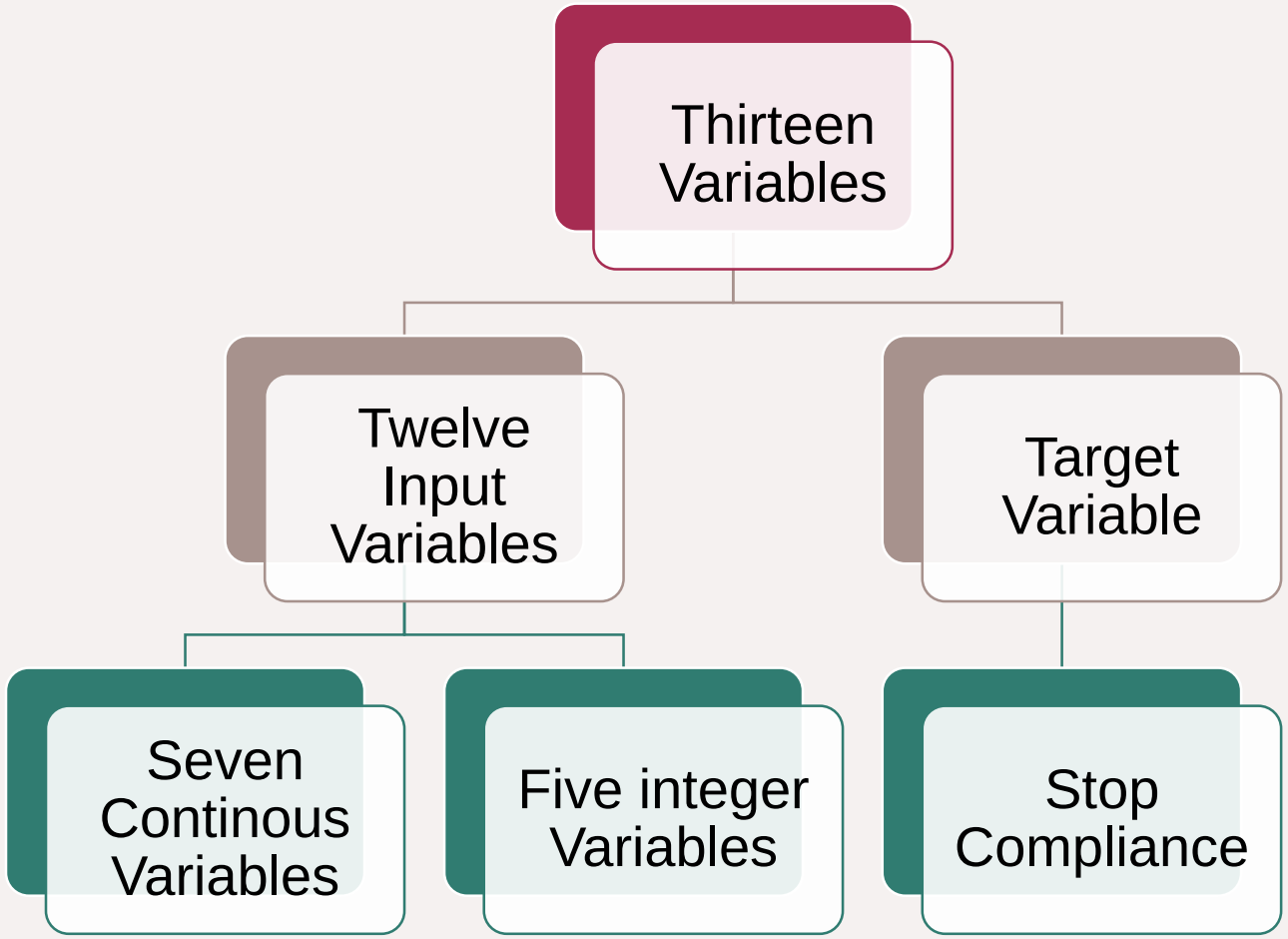


EDA on Barcelona Dataset





Input Variables and Target Variable for Quantitative Modelling



Binomial Logistic Regression Model

- Fits the **logistic function** $\boldsymbol{\sigma(z)}$ to model the relationship between input features and the predicted probability of the binary target being 1

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

where, $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n$

- If we consider, σ(z) = p i.e. the probability of stop_compliance being 1

  then, (1 - σ(z)) = (1-p) i.e the probability of stop_compliance being 0

- $\boldsymbol{log(odd\ ratio)} = ln\left(\frac{\sigma(z)}{1-\sigma(z)}\right)$

$$= ln\left(\frac{\frac{e^z}{e^z+1}}{1-\frac{e^z}{e^z+1}}\right) = ln(e^z) = \boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n} \qquad (3)$$

- **MLE** is used to learn **the optimal values of the coefficients** $\beta_0, \beta_1, .. \beta_n$ of the input features such that the model achieves **the best possible fit** to the observed dataset

Hypotheses on Expected Behaviours

H1: Stop compliance behaviour is **hindered by increased** population density and more trunk roads in the routes

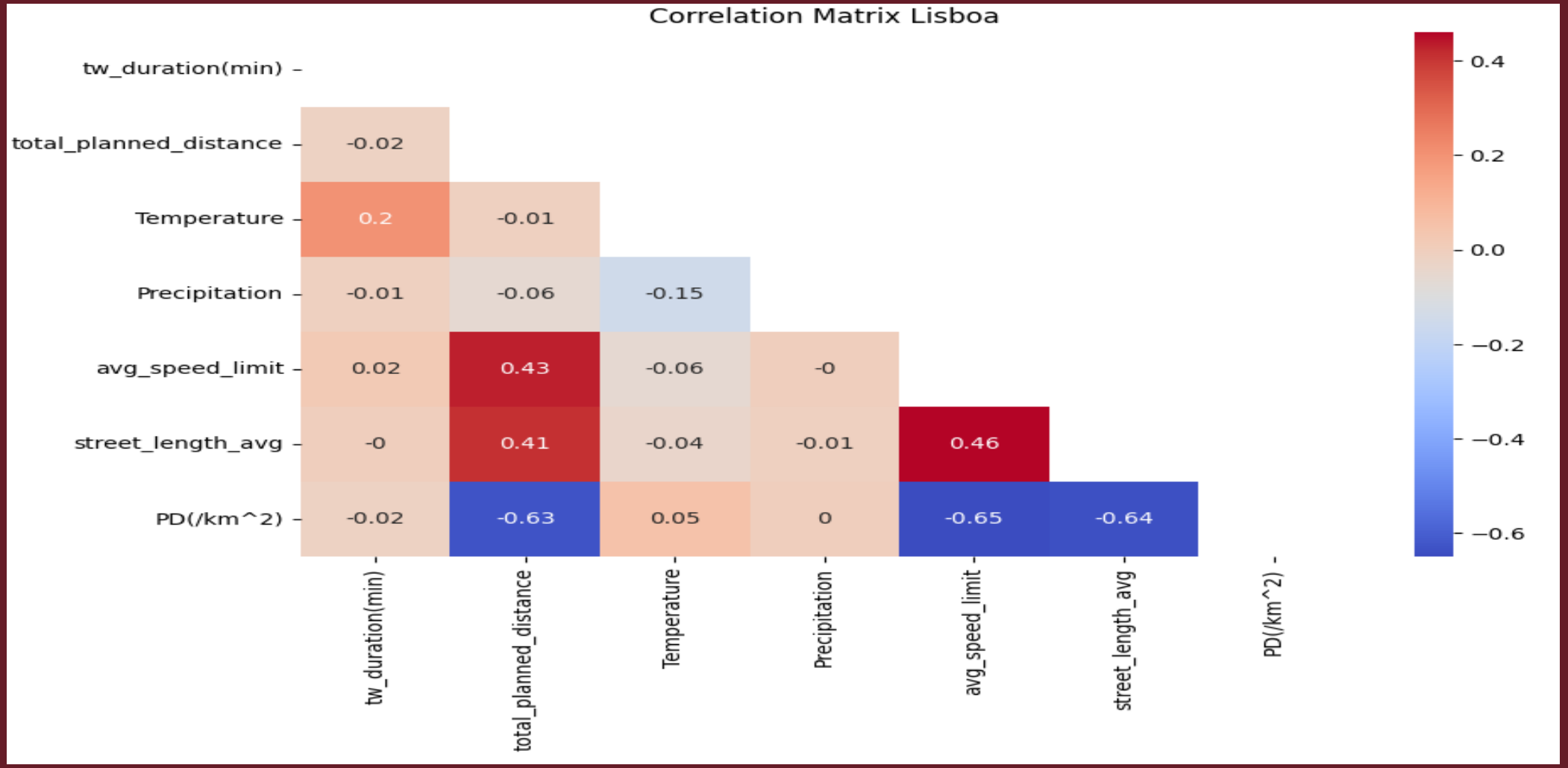
H2: The **shorter** the time window durations, the **lower** the odds of following the suggested stop orders

H3: As number of retailer company **grows** in the routes, driver stop compliance probability **drops**

H4: An **increase** in the number of delivery stops **reduces** the likelihood of compliance

H5: Offering deliveries in a **higher** number of different municipalities in a route **increases** the probability of compliance with suggested stop orders

Multicollinearity Check on Lisboa Dataset



Dummy Encoding of the Integer Variables

Categorical Variables	Intervals	Dummy Encoding			
trunk_category_low	[0,134]			1	0
trunk_category_high	[278,399]			0	1
municipality_category_low	[1,3]			1	0
municipality_category_high	[4,7]			0	1
stop_category_low	[10,34]	1	0	0	0
stop_category_medium	[35,59]	0	1	0	0
stop_category_med-high	[60,84]	0	0	1	0
stop_category_high	[85,100]	0	0	0	1
company_category_low	[1,5]		1	0	0
company_category_medium	[6,10]		0	1	0
company_category_high	[11,16]		0	0	1
intersection_category_low	[829,4238]		1	0	0
intersection_category_medium	[4239,7648]		0	1	0
intersection_category_high	[7649,11058]		0	0	1

Logistic Regression on Lisboa Dataset

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

- Time-window duration is the **most important factor**.
- The **second strongest factor** is the dummy variable trunk_category_high.
- The intersection count category turns out to be **the least important factor** in predicting compliance .

Factor Interpretation

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

The **negative coefficients** for the variables **PD(/km2) and trunk_category_high** reveal that stop compliance behaviour is **hindered** by increased population density and more trunk roads in the routes might be **due to higher traffic volumes and busy areas** . **Therefore, hypothesis-1 cannot be rejected.**

Factor Interpretation

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

One standard deviation **increase in time window duration** leads to a $(e^{0.272} - 1) * 100 \approx 31.3$ % **increase in the odds of stop compliance**. Therefore, as **the time window duration increases**, the probability of following the stops with suggested order also **increases. Thus , hypothesis-2 cannot be rejected**.

Factor Interpretation

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

Both medium and high category company counts reveal **a consistent linear decreasing effect** for compliance. Therefore, as the number of retailer companies **increases**, the probability of driver compliance **decreases**, **as supported by hypothesis-3.**

Factor Interpretation

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

- The stop_category_medium shows **no significant change** on compliance.
- Routes with moderately high number of stops displays **a relatively stronger negative** impact on compliance.
- But stop_category_high group shows **a smaller magnitude effect** than the stop_category_med.high group. Therefore, a **non-linear trend** is observed for stop count categories. **Thus, Hypothesis-4 would be rejected for stop count categories.**

Factor Interpretation

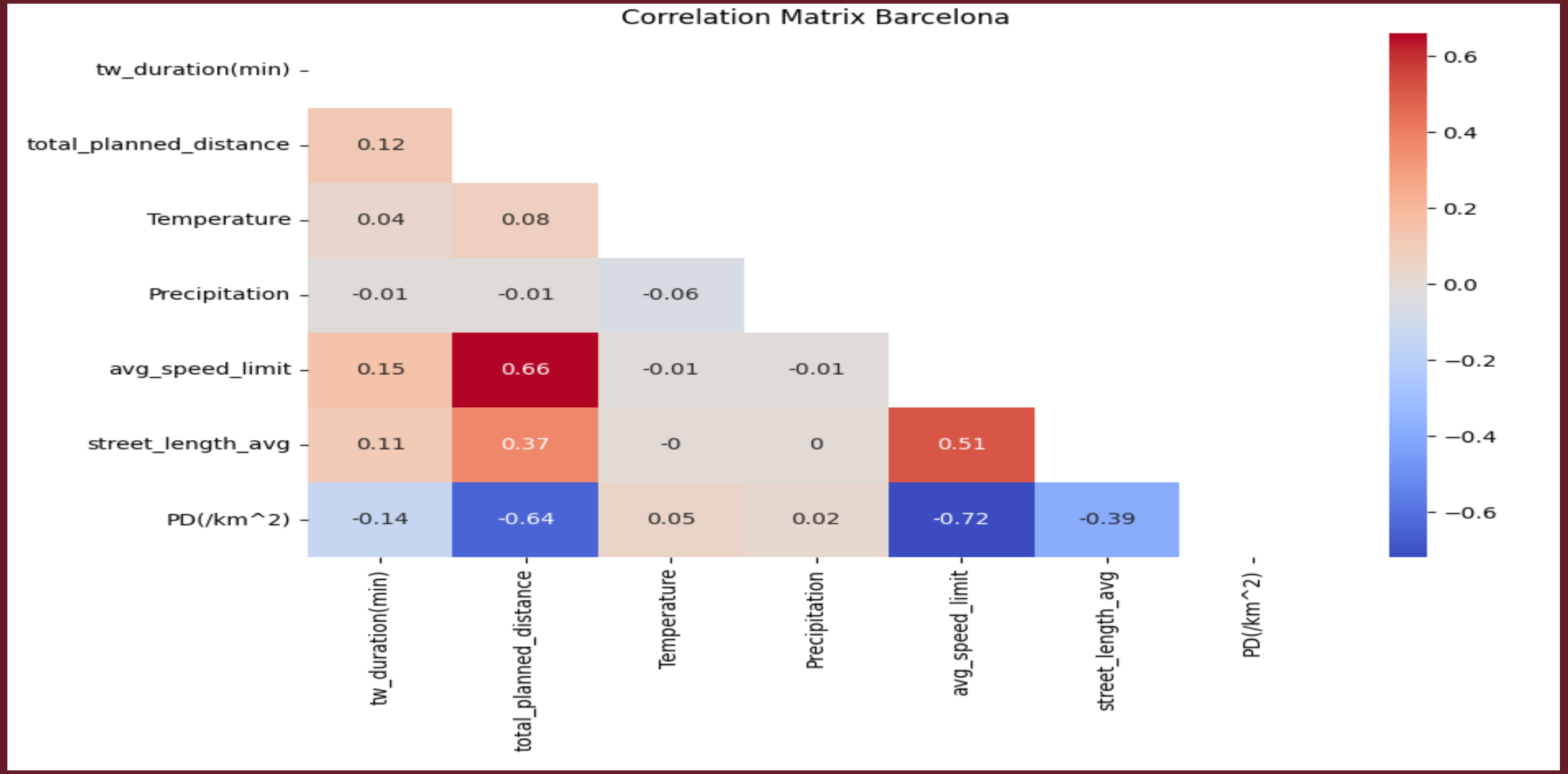
	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	-0.201*** (0.009)	-0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	-0.240*** (0.006)	-0.113*** (0.010)
Temperature	-0.061*** (0.004)	-0.048*** (0.004)
Precipitation	-0.097*** (0.004)	-0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		-0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		-0.013 (0.017)
stop_category_med.high		-0.105*** (0.017)
stop_category_high		-0.032*** (0.005)
company_category_medium		-0.053*** (0.004)
company_category_high		-0.103*** (0.004)
intersection_category_medium		-0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

The ‘municipality_category_high’ variable reveals that the stops of the routes offering deliveries in 4 to 7 different municipalities **are 2.3% more likely to be followed** than the stops of the routes offering deliveries in 1 to 3 different municipalities. **Therefore, drivers tend to follow the suggestion in the routes incorporating more unfamiliar places as expected from hypothesis-5.**

Factor Interpretation

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.270*** (0.003)	0.272*** (0.003)
PD..km.2.	−0.201*** (0.009)	−0.063*** (0.011)
avg_speed_limit	0.144*** (0.005)	0.030*** (0.009)
street_length_avg	−0.240*** (0.006)	−0.113*** (0.010)
Temperature	−0.061*** (0.004)	−0.048*** (0.004)
Precipitation	−0.097*** (0.004)	−0.074*** (0.004)
total_planned_distance	0.036*** (0.005)	0.027*** (0.005)
trunk_category_high		−0.187*** (0.011)
municipality_category_high		0.022*** (0.004)
stop_category_medium		−0.013 (0.017)
stop_category_med.high		−0.105*** (0.017)
stop_category_high		−0.032*** (0.005)
company_category_medium		−0.053*** (0.004)
company_category_high		−0.103*** (0.004)
intersection_category_medium		−0.014*** (0.004)
Constant	1.590*** (0.004)	1.597*** (0.004)

Multicollinearity Check on Barcelona Dataset



Dummy Encoding of the Integer Variables

Categorical Variables	Intervals	Dummy Encoding			
trunk_category_low	[0,42]			1	0
trunk_category_high	[95,123]			0	1
municipality_category_low	[1,4]			1	0
municipality_category_high	[5,9]			0	1
stop_category_low	[11,41]	1	0	0	0
stop_category_medium	[42,72]	0	1	0	0
stop_category_med-high	[73,103]	0	0	1	0
stop_category_high	[104,132]	0	0	0	1
company_category_low	[1,11]		1	0	0
company_category_medium	[12,22]		0	1	0
company_category_high	[23,34]		0	0	1
intersection_category_low	[9,3419]		1	0	0
intersection_category_medium	[3420,6830]		0	1	0
intersection_category_high	[6831,10240]		0	0	1

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

- Time window duration has **the least effect size** on compliance
- Precipitation **don't show any statistically significant effect** on changing compliance
- Population density has become **the second strongest factor**
- The **most significant factor** is the dummy variable company_category_high
- Intersection count category also **shows no statistically significant effect** on compliance

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

One standard deviation **increase in population density** leads to a $(e^{-0.070}-1) * 100 \approx 7\%$ **decrease in the odds of stop compliance**. The **negative coefficient** of **high trunk road category** variable also supports a **decreasing** effect in compliance. **Thus,hypothesis -1 can not be rejected**.

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
V1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

Though time window duration turns out to be the **least important factor** for Barcelona dataset, longer time window duration still **positively effects compliance, as expected from hypothesis-2.**

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

Drivers are about **6% less likely** to follow the suggestion when dealing with **medium numbers of company products** and nearly **8% less likely** to follow the suggestion when dealing with **high numbers of company products**. **Therefore, hypothesis-3 cannot be rejected**

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

Hypothesis-4 would be rejected for stop count categories due to non-linearity , as the **'stop_category_high'** shows **a lesser impact size** than other categories which leads to a non-linear trend

Logistic Regression on Barcelona Dataset

	y1 (1)	y2 (2)
v1		
tw_duration.min.	0.013*** (0.003)	0.009*** (0.003)
total_planned_distance	-0.065*** (0.004)	-0.055*** (0.004)
Temperature	-0.042*** (0.003)	-0.030*** (0.003)
Precipitation	-0.005* (0.003)	0.001 (0.003)
avg_speed_limit	0.058*** (0.005)	0.032*** (0.005)
street_length_avg	-0.003 (0.003)	0.013*** (0.004)
PD..km.2.	-0.123*** (0.005)	-0.070*** (0.006)
trunk_category_high		-0.048*** (0.005)
municipality_category_high		0.034*** (0.003)
stop_category_medium		-0.022*** (0.005)
stop_category_med.high		-0.031*** (0.005)
stop_category_high		-0.014*** (0.003)
company_category_medium		-0.062*** (0.003)
company_category_high		-0.078*** (0.004)
intersection_category_medium		0.001 (0.003)
Constant	-0.387*** (0.003)	-0.388*** (0.003)

Routes offering deliveries in **a high number of municipalities** have **a higher likelihood of compliance** compared to the routes with a **low number of municipalities**. **Therefore, hypothesis 5 cannot be rejected.**

Predictive Modelling

- The **classification problem** involves **predicting the likelihood** of a stop along a route being compliant
- The target variable, **"stop compliance"** is classified as **level 1** when a driver visits a stop in the suggested sequence order in undirected case and as **level 0** otherwise.
- Three commonly used algorithms are used for predicting classification: **logistic regression, random forest classifier, and gradient boosting classifier**
- Logistic regression is **a linear classifier** that assumes a **linear relationship** between the independent variables and the binary dependent variable. On the contrary, random forest and gradient boosting are **tree-based ensemble methods** that excel in capturing **non-linear relationships** between the dependent and independent variables

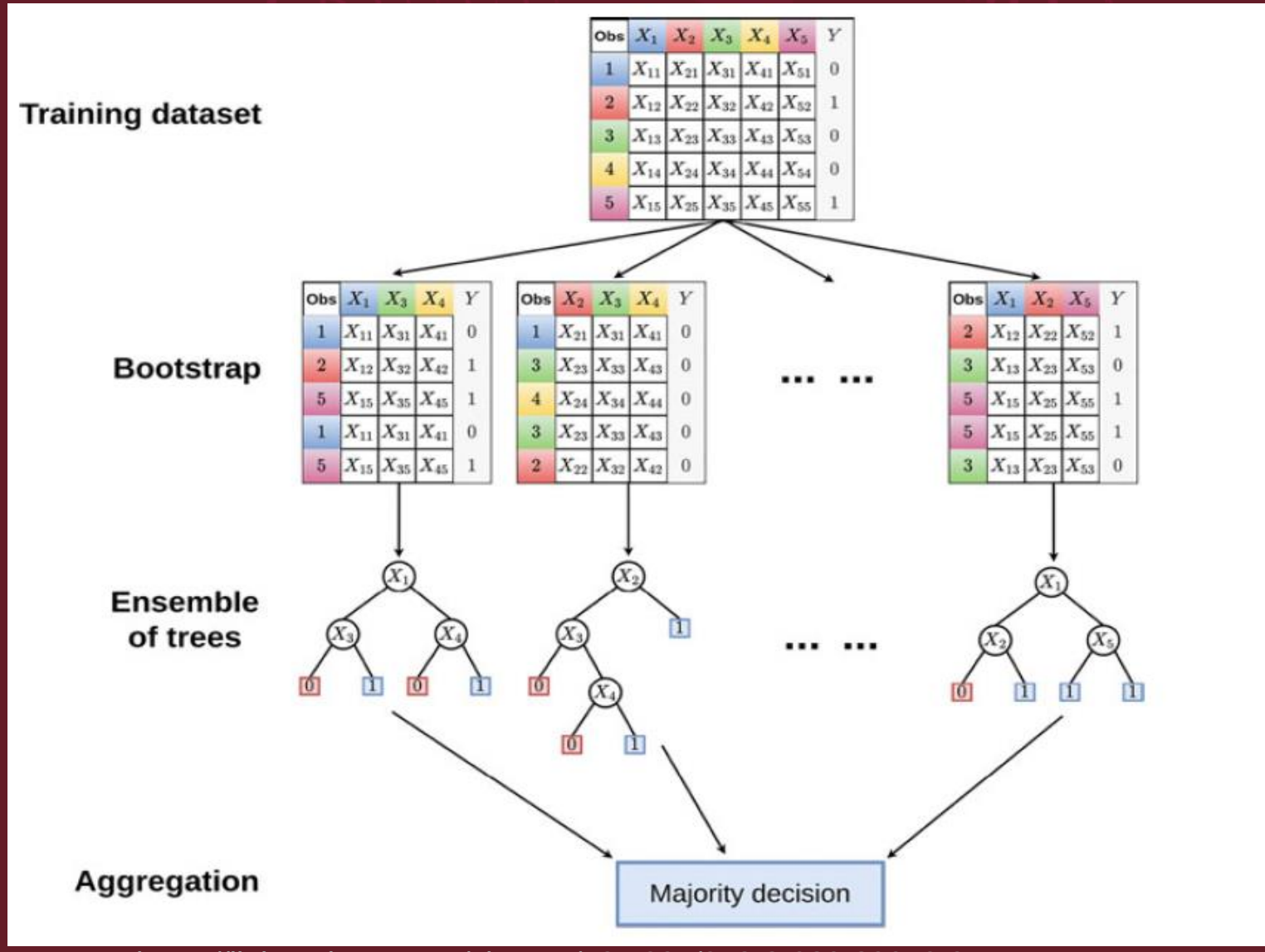
Random Forest Classifier

- Builds **an ensemble of decision trees simultaneously** using **bootstrapped** samples from the training data
- Each tree grows by **splitting nodes to minimize Gini impurity** calculated by

$$Gini(t) = 1 - \sum_{i=1}^{k} P_{i,t}^2$$

$$Gini(t_s) = \frac{|t_{right}|}{|t|} Gini(t_{right}) + \frac{|t_{left}|}{|t|} Gini(t_{left})$$

- **Final Prediction** is done by **Majority Voting** for decisions from all individual trees



https://link.springer.com/chapter/10.1007/978-3-030-89010-0_15

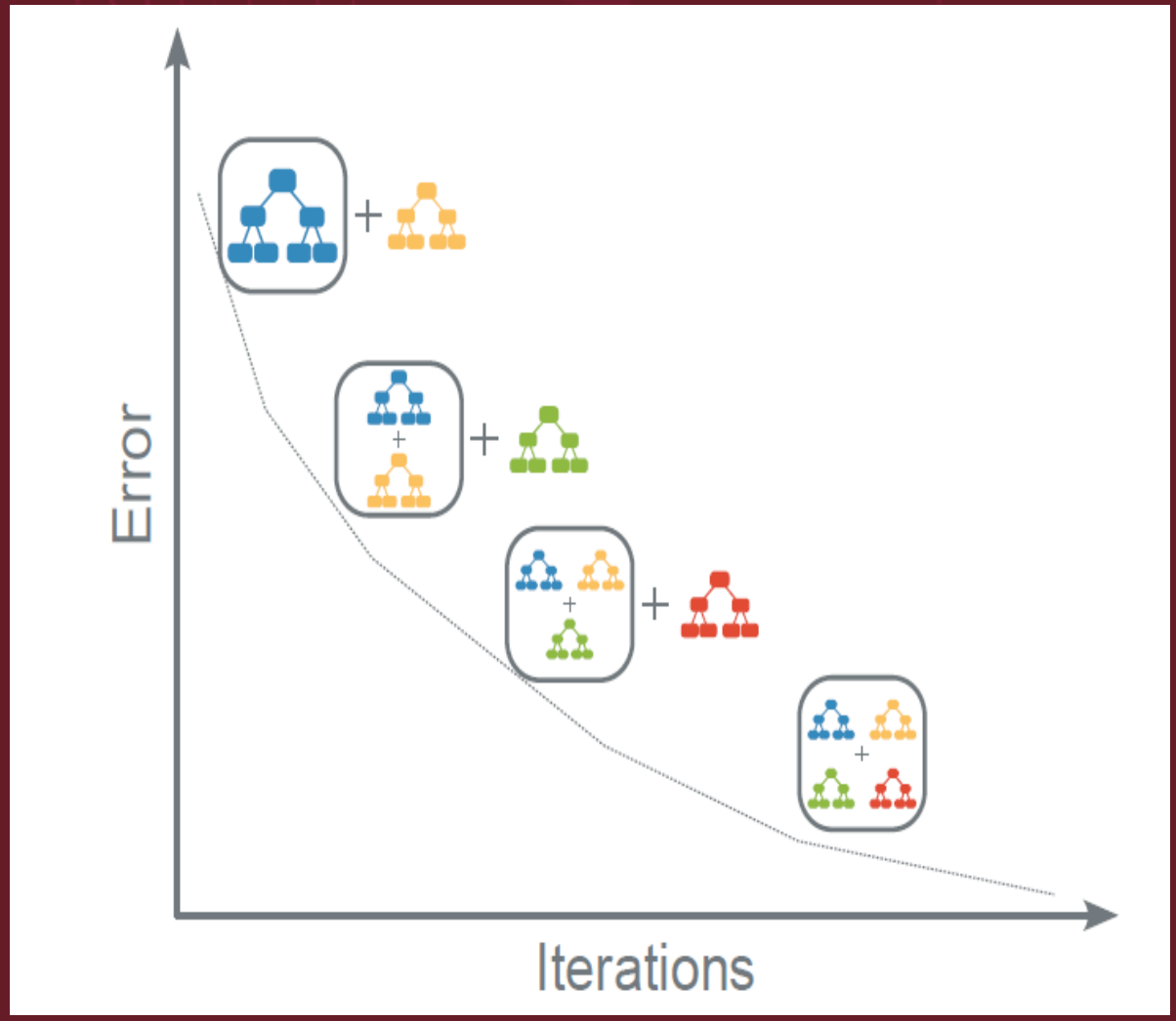
Gradient Boosting Classifier

- Constructs **an ensemble of decision trees sequentially**
- Starts with an **initial prediction** function F_o and **iteratively builds regression trees** h_m to correct the **errors of the preceding model's** predictions as

$$F_o(X_i) = log(\frac{P(Y=1)}{P(Y=0)})$$

$$F_m(X_i) = F_{m-1}(X_i) + \gamma_m * h_m(X_i)$$

- The **final prediction** is obtained by **summing the predictions of all decision trees** in the ensemble



https://www.researchgate.net/publication/368507910_Multi-teacher_knowledge_distillation_as_an_effective_method_for_compressing_ensembles_of_neural_networks

Performance Matrics

- **Specificity (true negative rate):** It measures the proportion of noncompliant stops correctly identified by a classifier.

$$Specificity = \frac{TN}{TN + FP}$$

- **Sensitivity (true positive rate):** It measures the proportion of compliant stops correctly identified by a classifier.

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

- **Accuracy:** It measures the **overall correctness** of a classifier by calculating the proportion of correct predictions.

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

Model Evaluation

To ensure a fair evaluation, we split each dataset into two parts: **80% for training and 20% for testing performance.**

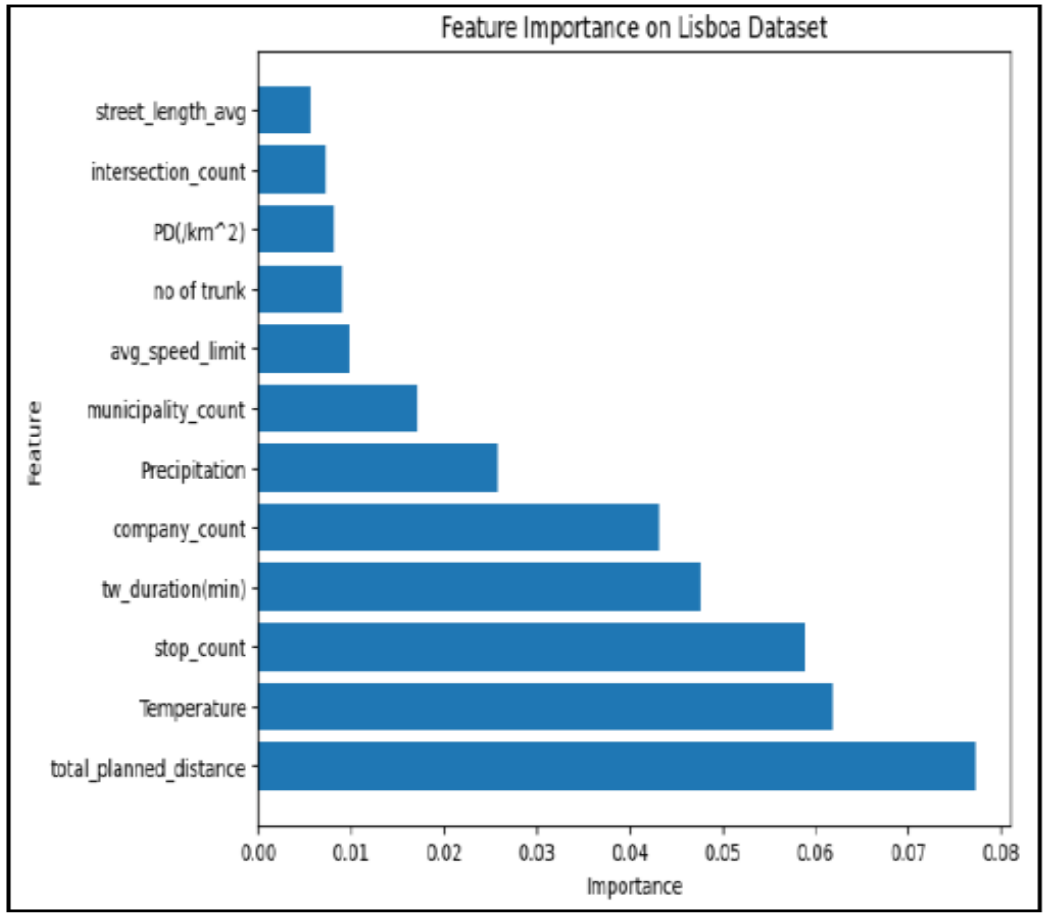
Model	Accuracy	Specificity	Sensitivity
Logistics Regression	0.59	0.56	0.60
Random Forest Classification	0.75	0.68	0.77
Gradient Boosting Classification	0.62	0.59	0.63

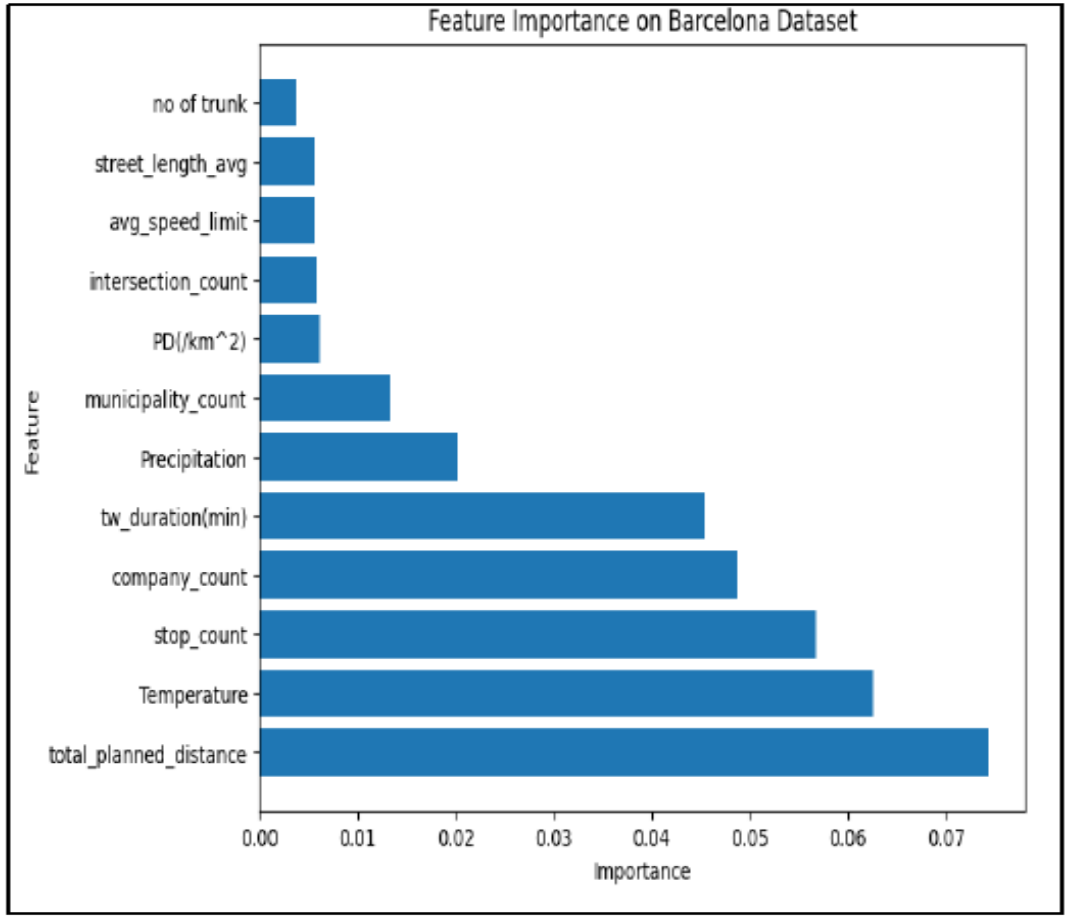
Table 5: Model Performance for Lisboa Dataset

Model	Accuracy	Specificity	Sensitivity
Logistics Regression	0.57	0.60	0.52
Random Forest Classification	0.73	0.79	0.64
Gradient Boosting Classification	0.64	0.70	0.56

Table-6: Model Performance for Barcelona Dataset

Feature Importance from Random Forest Classifier





Partial Dependence Plots

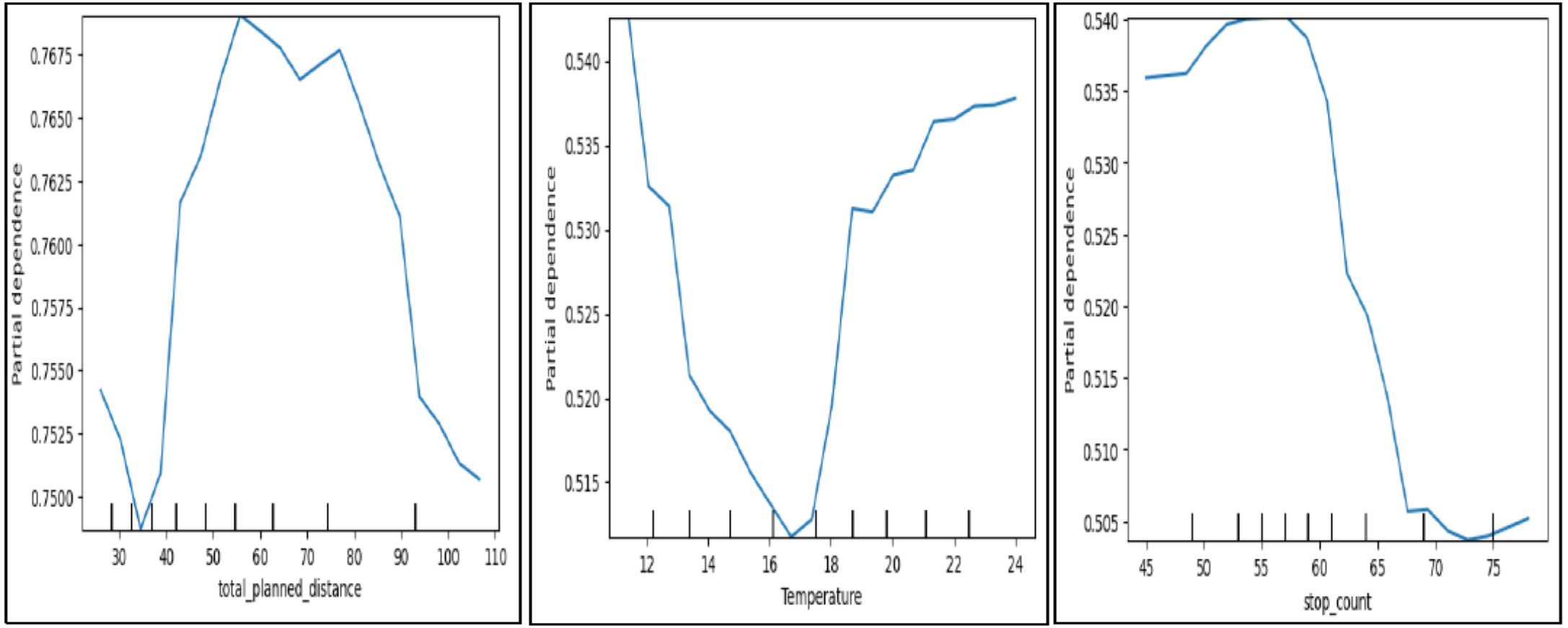


Figure-36: PDP for Random Forest Model (RF) from Lisboa Dataset

Partial Dependence Plots

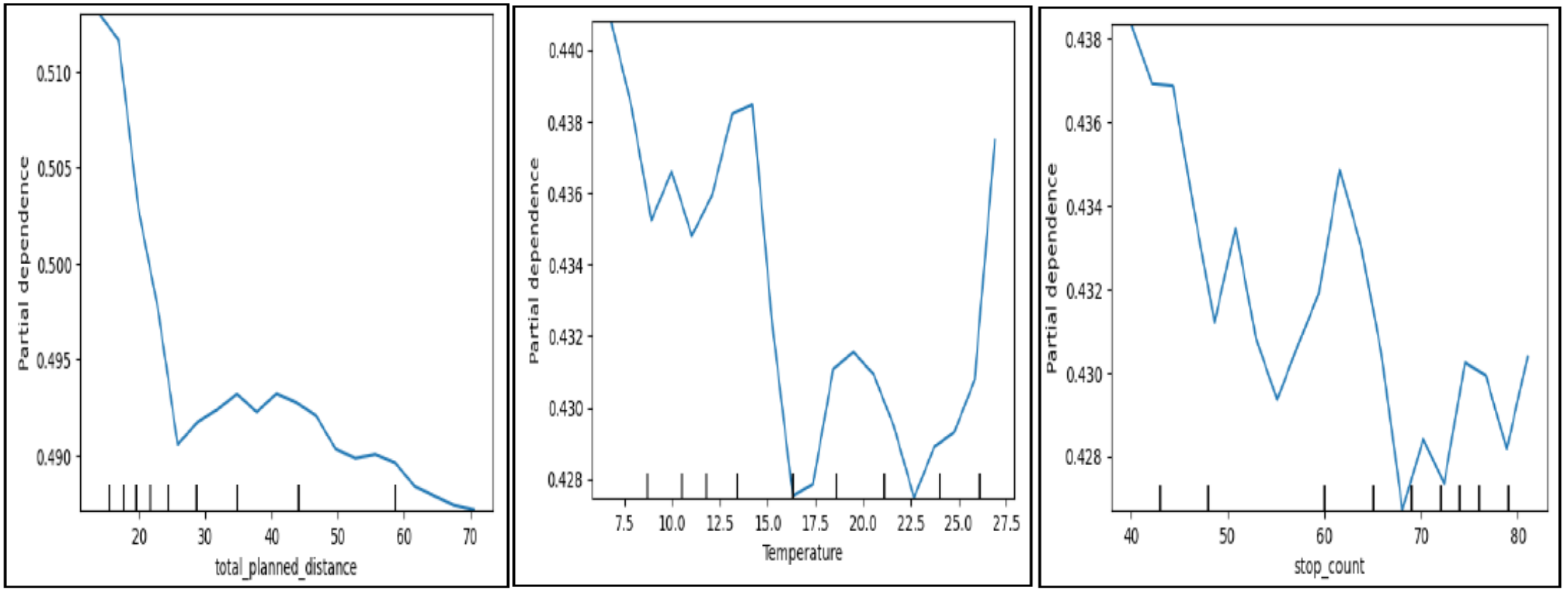


Figure-37: PDP for Random Forest Model (RF) from Barcelona Dataset

Conclusion and Future Work

- Real-life situations are **complex**, which makes it **challenging for drivers** to always follow **the fastest or shortest suggested routes** from conventional optimization tools.
- This study uncovered some valuable insights **into driver behaviour patterns** which will help for more **realistic route planning**.
- **Predicting driver's compliance behaviour** can be useful for companies to identify orders that might not be delivered as planned and **notify the customers accordingly**.
- We examined compliance for **single-stop observation** with certain assumptions where studying compliance **over longer stop sequences** will allow academics to expand on this analysis in the future.
- A more intuitive prediction can be made by adding **more relevant variables**, such as **drivers' personal characteristics and real-time information** on traffic, weather conditions, etc. However, it is difficult to **properly understand drivers' compliance patterns** due to subtle factors like **habits or personal concerns** that may affect their actions.

References

1. D. Le, ‘Two-stage route planning algorithm for last mile delivery’, masters, Concordia University, 2022.
2. Y. Dai et al., ‘Dynamic prediction of drivers’ personal routes through machine learning’, Dec. 2016, pp. 1–8.
3. ‘Last Mile Delivery Transportation Market Size, Report 2022-2030’. https://www.precedenceresearch.com/last-mile-delivery-transportation-market.
4. S. Srivatsa Srinivas and M. S. Gajanand, ‘Vehicle routing problem and driver behaviour: a review and framework for analysis’, Transp. Rev., vol. 37, no. 5, pp. 590–611, Sep. 2017.
5. T. Yamamoto, R. Kitamura, and J. Fujii, ‘Drivers’ Route Choice Behavior: Analysis by Data Mining Algorithms’, Transp. Res. Rec., vol. 1807, no. 1, pp. 59–66, Jan. 2002.
6. X. Lai, H. Fu, J. Li, and Z. Sha, ‘Understanding drivers’ route choice behaviours in the urban network with machine learning models’, IET Intell. Transp. Syst., vol. 13, no. 3, pp. 427–434, 2019.

References

7. G. Boeing, ‘OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks’, Comput. Environ. Urban Syst., vol. 65, pp. 126–139, Sep. 2017.

8. ‘Python Library | Meteostat Developers’. https://dev.meteostat.net/python/#installation (accessed Jul. 01, 2023).

9. ‘Europe: Population Statistics in Maps and Charts for Cities, Agglomerations and Administrative Divisions of all Countries in Europe’. https://www.citypopulation.de/Europe.html (accessed Jul. 31, 2023).

10. F. Pedregosa et al., ‘Scikit-learn: Machine Learning in Python’, J. Mach. Learn. Res., vol. 12, no. 85, pp. 2825–2830, 2011.

11. G. Biau and E. Scornet, ‘A random forest guided tour’, TEST, vol. 25, no. 2, pp. 197–227, Jun. 2016.

12. C. Molnar, Interpretable Machine Learning. Lulu.com, 2020.