

Univerzita Hradec Králové  
Fakulta informatiky a managementu

## **Semestrální projekt BIN**

Bc. Adéla Leppeltová

2024/25

# Obsah

<b>Nastavení cílů .....</b>	<b>3</b>
<b>Popis dat .....</b>	<b>3</b>
<i>Základní charakteristiky dat.....</i>	<i>4</i>
<b>Metodika procesu .....</b>	<b>5</b>
<i>Úpravy dat.....</i>	<i>5</i>
<i>Výběr modelu .....</i>	<i>5</i>
<b>Vyhodnocení výstupů .....</b>	<b>6</b>
<i>Model lineární regrese .....</i>	<i>6</i>
<i>Model C&amp;R Tree .....</i>	<i>7</i>
<i>Model Random Forest .....</i>	<i>7</i>
<i>Model Neural Net.....</i>	<i>7</i>
<i>Kombinovaný model.....</i>	<i>8</i>
<b>Vyhodnocení modelu vzhledem k cílům .....</b>	<b>8</b>
<b>Možná omezení a jejich řešení .....</b>	<b>8</b>
<b>Zdroje .....</b>	<b>9</b>

## Nastavení cílů

Cílem je provést predikci počtu půjčených kol na základě enviromentálních (počasí, teplota, pocitová teplota, vlhkost vzduchu a vítr) a časových proměnných (rok, roční období, měsíc, den v týdnu, pracovní den/víkend).

## Popis dat

Dataset obsahuje dva soubory – hour.csv, který obsahuje počty vypůjčených kol za hodinu a day.csv, který obsahuje počty vypůjčených kol za den. Oba soubory dále obsahují enviromentální a časové proměnné. Pro práci byl vybrán pouze soubor hour.csv.

Odkaz na data: <https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset>

Oba soubory (hour.csv, day.csv) obsahují následující sloupce, kromě atributu *hr*, který se nenachází v souboru day.csv,

**instant**: index záznamu

**dteday** : datum

**season** : roční období (1:jaro, 2:léto, 3:podzim, 4:zima)

**yr** : rok (0: 2011, 1:2012)

**mnth** : měsíc (od 1 do 12)

**hr** : hodina (od 0 do 23)

**holiday** : údaj, zda jsou prázdniny, nebo ne

**weekday** : den v týdnu

**workingday** : pokud den není víkend ani svátek, je hodnota 1, jinak je 0

**weathersit** : údaj o počasí

- 1: Jasno, částečně zataženo
- 2: Mlha + zataženo, mlha + roztrhaná oblačnost, mlha + pár mraků, mlha
- 3: Slabé sněžení, slabý déšť + bouřka + mlha, sněžení + mlha
- 4: Silný déšť + kroupy + bouřka + mlha, sněžení + mlha

**temp** : normalizovaná teplota ve stupních Celsia, hodnoty jsou rozděleny na 41

**atemp**: normalizovaná pocitová teplota ve stupních Celsia, hodnoty jsou rozděleny na 50

**hum**: normalizovaná vlhkost vzduchu, hodnoty jsou rozděleny na 100

**windspeed**: normalizovaná rychlost větru, hodnoty jsou rozděleny na 67

**casual**: počet občasných (neregistrovaných) uživatelů

**registered**: počet registrovaných uživatelů

**cnt**: celkový počet vypůjčených kol (registrovanými i neregistrovanými uživateli) (1)

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1		1 2011-01-01	1	0	1	0	0	6	0	1	0.240	0.288	0.810	0.000	3	13	16
2		2 2011-01-01	1	0	1	1	0	6	0	1	0.220	0.273	0.800	0.000	8	32	40
3		3 2011-01-01	1	0	1	2	0	6	0	1	0.220	0.273	0.800	0.000	5	27	32
4		4 2011-01-01	1	0	1	3	0	6	0	1	0.240	0.288	0.750	0.000	3	10	13
5		5 2011-01-01	1	0	1	4	0	6	0	1	0.240	0.288	0.750	0.000	0	1	1
6		6 2011-01-01	1	0	1	5	0	6	0	2	0.240	0.258	0.750	0.090	0	1	1
7		7 2011-01-01	1	0	1	6	0	6	0	1	0.220	0.273	0.800	0.000	2	0	2
8		8 2011-01-01	1	0	1	7	0	6	0	1	0.200	0.258	0.860	0.000	1	2	3
9		9 2011-01-01	1	0	1	8	0	6	0	1	0.240	0.288	0.750	0.000	1	7	8
10		10 2011-01-01	1	0	1	9	0	6	0	1	0.320	0.348	0.760	0.000	8	6	14



































Obrázek 1: Náhled dat

## Základní charakteristiky dat

Celkový počet záznamů v souboru hour.csv je 17 379. Soubor neobsahuje chybějící ani nulové hodnoty. Některé proměnné mají odlehlé hodnoty, jejich konkrétní počty jsou vyobrazeny v obrázku 3.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
instant		Continuous	1	17379	8690	5017.029	0	--	17379
dteday		Continuous	2011-01-01	2012-12-31	--	--	--	--	17379
season		Continuous	1	4	2.502	1.107	-0.005	--	17379
yr		Continuous	0	1	0.503	0.500	-0.010	--	17379
mnth		Continuous	1	12	6.538	3.439	-0.009	--	17379
hr		Continuous	0	23	11.547	6.914	-0.011	--	17379
holiday		Continuous	0	1	0.029	0.167	5.639	--	17379
weekday		Continuous	0	6	3.004	2.006	-0.003	--	17379
workingday		Continuous	0	1	0.683	0.465	-0.785	--	17379
weathersit		Continuous	1	4	1.425	0.639	1.228	--	17379
temp		Continuous	0.020	1.000	0.497	0.193	-0.006	--	17379
atemp		Continuous	0.000	1.000	0.476	0.172	-0.090	--	17379
hum		Continuous	0.000	1.000	0.627	0.193	-0.111	--	17379
windspeed		Continuous	0.000	0.851	0.190	0.122	0.575	--	17379
casual		Continuous	0	367	35.676	49.305	2.499	--	17379
registered		Continuous	0	886	153.787	151.357	1.558	--	17379
cnt		Continuous	1	977	189.463	181.388	1.277	--	17379

Obrázek 2: Základní statistiky souboru hour.csv

Complete fields (%): 100 %		Complete records (%): 100 %										
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
 instant	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 dteday	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 season	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 yr	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 mnth	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 hr	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 holiday	 Continuous	0	500	None	Never	Fixed	100	17379	0	0	0	0
 weekday	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 workingday	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 weathersit	 Continuous	3	0	None	Never	Fixed	100	17379	0	0	0	0
 temp	 Continuous	0	0	None	Never	Fixed	100	17379	0	0	0	0
 atemp	 Continuous	1	0	None	Never	Fixed	100	17379	0	0	0	0
 hum	 Continuous	22	0	None	Never	Fixed	100	17379	0	0	0	0
 windspeed	 Continuous	102	5	None	Never	Fixed	100	17379	0	0	0	0
 casual	 Continuous	411	56	None	Never	Fixed	100	17379	0	0	0	0
 registered	 Continuous	371	0	None	Never	Fixed	100	17379	0	0	0	0
 cnt	 Continuous	244	0	None	Never	Fixed	100	17379	0	0	0	0

Obrázek 3: Základní charakteristiky neupravených dat

# Metodika procesu

## Úpravy dat

Po načtení dat byly data upraveny pomocí uzlu „Type“. Aby nedošlo k chybné interpretaci proměnných byly proměnné *season*, *weekday* a *weathersit* převedeny na nominální typ. A proměnné *yr*, *holiday* a *workingday* byly převedeny na typ „Flag“, obsahují pouze hodnoty 0 nebo 1.

Proměnné *temp*, *atemp*, *hum* a *windspeed* jsou již normalizovány, úpravy nebyly třeba. Ostatní proměnné byly ponechány v původním stavu.

V tomto uzlu byla také určena cílová proměnná *cnt*.

Field	Sample Graph	Measurement	Min	Max	Mean	Correlation	Correlation T	Correlation T df.	Correlation T sig.	Covarian...	Std. Dev	Skewness	Unique	Valid
dteday		Continuous	2011-01-01	2012-12-31	--	--	--	--	--	--	--	--	--	17379
season		Nominal	1	4	--	--	--	--	--	--	--	--	4	17379
yr		Flag	0	1	--	--	--	--	--	--	--	--	2	17379
mnth		Continuous	1	12	6.538	0.121	16.020	17377.000	0.000	75.248	3.439	-0.009	--	17379
hr		Continuous	0	23	11.547	0.394	56.521	17377.000	0.000	494.239	6.914	-0.011	--	17379
holiday		Flag	0	1	--	--	--	--	--	--	--	--	2	17379
weekday		Nominal	0	6	--	--	--	--	--	--	--	--	7	17379
workingday		Flag	0	1	--	--	--	--	--	--	--	--	2	17379
weathersit		Nominal	1	4	--	--	--	--	--	--	--	--	4	17379
temp		Continuous	0.020	1.000	0.497	0.405	58.352	17377.000	0.000	14.138	0.193	-0.006	--	17379
atemp		Continuous	0.000	1.000	0.476	0.401	57.691	17377.000	0.000	12.498	0.172	-0.090	--	17379
hum		Continuous	0.000	1.000	0.627	-0.323	-44.976	17377.000	0.000	-11.300	0.193	-0.111	--	17379
windspeed		Continuous	0.000	0.851	0.190	0.093	12.344	17377.000	0.000	2.069	0.122	0.575	--	17379
casual		Continuous	0	367	35.676	0.695	127.265	17377.000	0.000	6211.710	49.305	2.499	--	17379
registered		Continuous	0	886	153.787	0.972	546.820	17377.000	0.000	26689.7...	151.357	1.558	--	17379
cnt		Continuous	1	977	189.463	--	--	--	--	--	181.388	1.277	--	17379

Obrázek 4: Základní statistiky upravených dat

## Výběr modelu

Před výběrem modelu byly data rozděleny na trénovací a testovací část v poměru 70:30. Dále byl proveden výběr příznaků (feature selection), kde jako cílová proměnná (target) byl zvolen celkový počet vypůjčených kol (proměnná *cnt*) a jako vstupní proměnné byly použity všechny zbylé proměnné (*season*, *yr*, *mnth*, *hr*, *holiday*, *weekday*, *workingday*, *weathersit*, *temp*, *atemp*, *hum*, *windspeed*) kromě těch, které přímo souvisely s cílovou proměnnou (*casual*, *registered*) a proměnné *instant* a *dteday*. Výstup je vyobrazen na následujícím obrázku.



## Model C&R Tree

Model C&R Tree na trénovací sadě dosáhl korelace 0,802 a MAE 72,24. Na testovací sadě dosáhl korelace 0,807 a MAE 71,964. V porovnání s ostatními modely je tento průměrný.

Comparing \$R-cnt with cnt		
'Partition'	1_Training	2_Testing
Minimum Error	-484,712	-484,712
Maximum Error	506,322	467,322
Mean Error	-0,361	-1,613
Mean Absolute Error	72,24	71,964
Standard Deviation	108,222	107,361
Linear Correlation	0,802	0,807
Occurrences	12 204	5 175

Obrázek 7: Výsledky modelu C&R Tree

## Model Random Forest

Tento model se v porovnání s ostatními jeví jako nejúspěšnější. Na trénovací sadě dosáhl korelace 0,996 a MAE 9,872. Na testovací sadě dosáhl korelace 0,973 a MAE 25,523.

Comparing \$RL-cnt with cnt		
'Partition'	1_Training	2_Testing
Minimum Error	-151,35	-513,78
Maximum Error	174,87	309,08
Mean Error	-0,488	-1,413
Mean Absolute Error	9,872	25,523
Standard Deviation	16,532	41,856
Linear Correlation	0,996	0,973
Occurrences	12 204	5 175

Obrázek 8: Výsledky modelu Random Forest

## Model Neural Net

Model Neural Net na trénovací sadě dosáhl korelace 0,904 a MAE 53,745. Na testovací sadě dosáhl korelace 0,906 a MAE 54,023. V porovnání s ostatními je model Neural Net třetí nejúspěšnější.

Comparing \$N-cnt with cnt		
'Partition'	1_Training	2_Testing
Minimum Error	-385,608	-368,073
Maximum Error	500,487	401,294
Mean Error	-2,502	-4,186
Mean Absolute Error	53,745	54,023
Standard Deviation	77,358	76,888
Linear Correlation	0,904	0,906
Occurrences	12 204	5 175

Obrázek 9: Výsledky modelu Neural Net

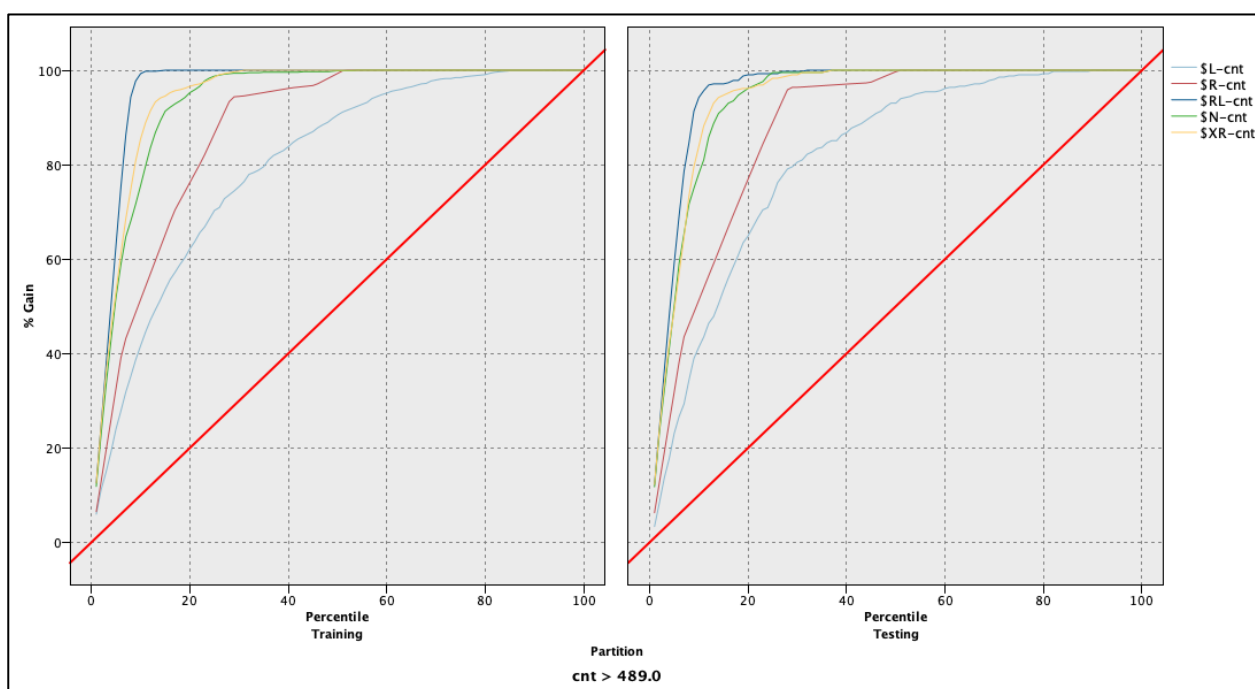
## Kombinovaný model

Kombinovaný model ve srovnání s ostatními se jeví jako druhý nejúspěšnější, na trénovací sadě dosáhl korelace 0,932 a MAE 50,451. Na testovací sadě dosáhl korelace 0,925 a MAE 52,659.

Comparing \$XR-cnt with cnt		
'Partition'	1_Training	2_Testing
Minimum Error	-270,651	-308,41
Maximum Error	399,054	387,346
Mean Error	-0,838	-1,875
Mean Absolute Error	50,451	52,659
Standard Deviation	72,757	75,299
Linear Correlation	0,932	0,925
Occurrences	12 204	5 175

Obrázek 10: Výsledek kombinovaného modelu

V grafickém porovnání všech modelů je patrné, že model Random Forest je nejúspěšnější., naopak model lineární regrese si vedl nejhůře.



Obrázek 11: Grafické porovnání modelů

## Vyhodnocení modelu vzhledem k cílům

Modely úspěšně predikují počty půjčených kol na základě zvolených proměnných. Nejúspěšnějšími byl model Random Forest a kombinovaný model, oba dosáhly vysoké úspěšnosti. Lze tedy stanovit, že cíl práce byl naplněn.

## Možná omezení a jejich řešení

Jedním z možných omezení je, že není zohledněna situace, kdy kola nebyla ve stanicích dostupná. Nižší počty půjčených kol nemusí odrážet skutečnou poptávku. Přidání dat



o dostupnosti kol v jednotlivých stanicích by mohlo zlepšit přesnost predikce a dále také přispět k optimalizaci rozmístění kol.

Další zajímavé poznatky by mohlo přinést přidání proměnných sledujících jednotlivá kola (například počet vypůjčení, ujetá vzdálenost, průměrná rychlost). Pomocí těchto informací by bylo možné predikovat servis kola.

## **Zdroje**

1. **Fanaee-T, H.** Bike Sharing [Dataset]. *UCI Machine Learning Repository*. [Online] 2013. <https://doi.org/10.24432/C5W894>.