Seoul Bike Sharing Demand Data Set

Daniel Rodriguez, Jorge Diaz,
Departamento de Ingeniería de Sistemas, Universidad de Antioquia
Medellín, Colombia
daniel.rodriguez1@udea.edu.co
jelias.diaz@udea.edu.co

***Abstract*— Las actividades propuestas para el desarrollo del proyecto buscan que cada uno de los grupos de estudiantes presenten todo el diseño, análisis y simulación de un sistema de predicción basado en técnicas de aprendizaje de máquina; describiendo el problema y su contexto en términos del estado del arte, especificando cada una de las etapas del desarrollo del trabajo, los modelos con sus respectivas restricciones, la metodología de validación, los resultados de las simulaciones y las conclusiones obtenidas.**

***Index Terms*—Modelos fenomenológicos, aprendizaje de máquina.**

INTRODUCCIÓN

El alquiler de bicicletas públicas es el nuevo sistema de transporte público que prospera en el mundo, la bicicleta pública no solo en los viajes de corta distancia puede aprovechar las ventajas flexibles y eficientes, sino que también puede extender de manera efectiva el alcance del servicio de transporte público. El tiempo real para dominar el número de bicicletas de alquiler de bicicletas para orientar las necesidades del público, es propicio para que los departamentos de formación y planificación desarrollen la política de alquiler. Sin embargo, existen pocos estudios sobre el impacto de los factores climáticos en el número de alquiler de bicicletas públicas hasta ahora. En este documento, proporcionamos un pronóstico para la demanda de alquiler de bicicletas en seoul, que se basa en los datos históricos de los datos de alquiler de bicicletas. De acuerdo con las características de los datos, utilizamos métodos de machine learning para pronosticar la demanda de alquiler de bicicletas.

I. PROBLEMA

Actualmente, las bicicletas de alquiler se introducen en muchas ciudades para mejorar la comodidad en la movilidad y son bien aceptadas como una manera de promover una actividad saludable a los ciudadanos. Para que este servicio sea efectivo es necesario que esté disponible en los momentos de mayor necesidad con el menor tiempo de espera. Por lo que se convierte en una prioridad el determinar en qué horarios y condiciones se hace el mayor uso del servicio de alquiler para poder asignar las bicicletas necesarias para disminuir el tiempo de espera.

El dataset está compuesto de 8465 filas en las que el servicio está disponible, con 14 variables y sin datos nulos, de las cuales 408 representan días festivos, lo que deja un total de 8057 que representan los días de la semana que no son festivos y el sistema está disponible. En lo que respecta a las variables categóricas: Se añadió el dia de la semana que representa la fecha, se Determinó dia de la semana de acuerdo a la fecha, adicionalmente se filtraron los datos nulos o vacíos (Días no funcionales), finalmente se dividió el dataset en los días festivos y normales.

	Variable	Tipo	Clase	Valores	Formato
1	Date	Date	Categórico	1-12-2017 - 30-11-2018	dd-mm-aaaa
2	Rented Bike count	Integer	Continuo	0 - 3556	
3	Hour	Integer	Continuo	0 - 23	
4	Temperature	Double	Continuo	-17.8 - 39.4	°C
5	Humidity	Integer	Continuo	0 - 98	%
6	Wind speed	Integer	Continuo	0 - 74	m/s
7	Visibility	Integer	Continuo	27 - 2000	10m
8	Dew point temperature	Double	Continuo	-30.6 - 27.2	°C
9	Solar Radiation	Integer	Continuo	0 - 352	MJ/m2
10	Rainfall	Integer	Continuo	0 - 295	mm
11	Snowfall	Integer	Continuo	0 - 88	mm
12	Seasons	String	Categórico	Winter, Spring, Summer, Autumn	
13	Holiday	Boolean	Categórico	Holiday, No Holiday	
14	Functioning day	Boolean	Categórico	Yes, No	

Tabla 1: Variables de datos y descripción.

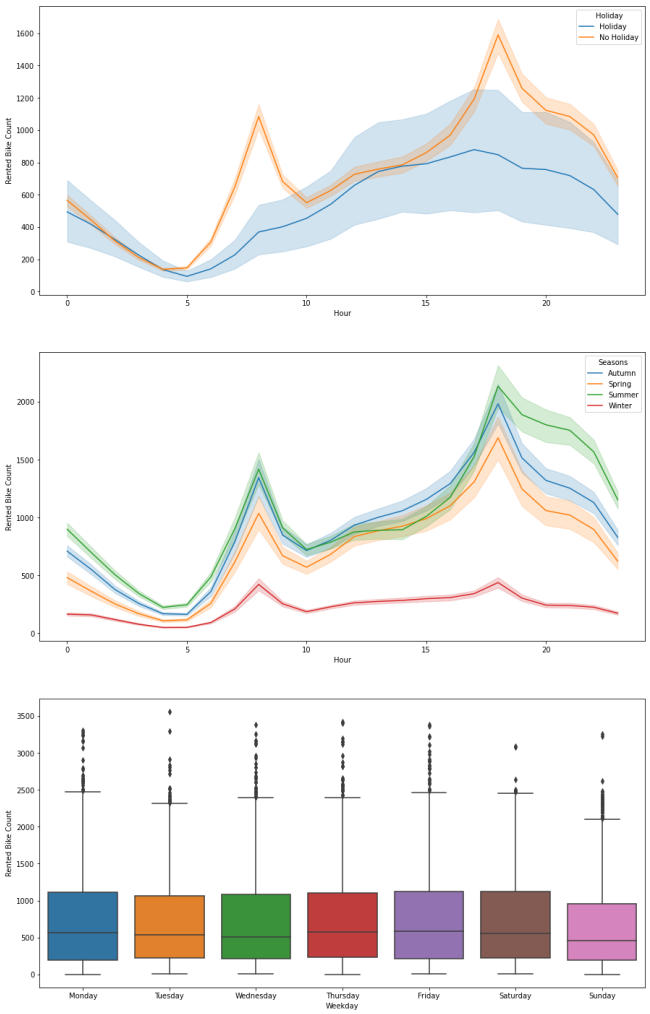


Figura 1: Gráfico de línea(Hour vs Rented Bike Count) y detalles del boxplot.

II. TRABAJOS RELACIONADOS

En la tabla 2: Se puede observar el resumen con la información más relevante para nuestros propósitos acerca de artículos que han trabajado con el mismo dataset y otro que no pero que también aborda el problema del alquiler de bicicletas.

	Modelos	Validación
[1]	- Cubist - Regularized Random Forest (RRF - Classification and regression trees - K-nearest neighbours (KNN) - Conditional inference tree	K-Fold Cross Validation (10 Folds - 3 Repeats) Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) R2 Coefficient of Variation (CV).
[2]	- Classification and regression trees - Random forest - Support vector machine - k-Nearest neighbors (KNN)	K-Fold Cross Validation (10 Fold - 3 Repeats) Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) R2
[3]	- Linear regression - Gradient Boosting Machine - Extreme Gradient Boosting Tree - Support Vector Machine - Boosted trees	K-Fold Cross Validation (10 Fold) Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) R2 Coefficient of Variation (CV).
[4]	- Regresión lineal múltiple - Random forests	R2 R2 Adjusted Sth.Error of the Estimate

Tabla 2: Detalles de los artículos investigados.

Los resultados del estudio [1]: Demuestran que el método más efectivo es el CUBIST, ya que se busca el mayor R^2 con los menores RMSE, MAE y CV.

			Training				Testing		
Models	Hyperparameter	R^2	RMSE	MAE	CV	R^2	RMSE	MAE	CV
CUBIST	*committees = 41* *neighbours = 3*	0.98	70.76	40.59	10.04	0.95	139.64	78.45	19.81
RRF	*mtry = 14* *coefReg = 0.505*	0.98	75.83	44.80	10.76	0.93	164.85	99.05	23.39
CART	*cp = 0.0001*	0.92	177.00	113.45	25.12	0.87	228.94	141.37	32.49
KNN	*k = 3*	0.89	213.11	128.74	30.24	0.77	299.88	188.94	42.55
CIT	*maxdepth = 19* *mincriterion = 0.01*	0.88	214.52	127.29	30.44	0.83	257.13	155.30	36.49

Figura 2: Detalle de los resultados del artículo uno.

En el estudio [2]: Los resultados demuestran que el método más efectivo es el RF, seguido por el SVM, con los mejores resultados en negrilla, cabe destacar que en este artículo la variable de "Date" se usa para determinar los días de la semana, los cuales finalmente son usados en su lugar. Además se usan para separar las muestras por temporadas.

Models	***R²***	**RMSE**	**MAE**	**CV**	**PI**
Training					
RF	**0.97**	**91.85**	**60.03**	**12.53**	**0.53**
SVM	0.93	159.80	90.04	22.68	0.76
CART	0.92	177.00	113.45	25.12	0.86
*k*NN	0.89	213.117	128.74	30.24	1
Testing					
RF	**0.88**	**216.01**	**130.52**	**30.63**	**0.73**
SVM	0.85	242.89	153.32	34.47	0.82
CART	0.87	228.94	141.37	32.49	0.78
*k*NN	0.77	299.88	188.94	42.55	1

Figura 3: Detalle de los resultados del artículo dos.

Finalmente en el estudio [3]: Los resultados demuestran que el modelo con mejor rendimiento es el Gradient Boosting Machine.

El estudio[4]: No se tuvieron en cuenta los resultados dado que no trabajaron con el mismo dataset y el enfoque del estudio es diferente. Sin embargo, abordan una problemática similar.

Models	Training				Testing			
	R^2	RMSE	MAE	CV (%)	R^2	RMSE	MAE	CV (%)
LM	0.55	431.72	321.84	61.15	0.55	427.71	322.32	61.03
GBM	0.96	117.81	79.77	16.68	0.92	172.73	109.78	24.64
SVM	0.92	173.58	96.32	24.58	0.85	241.94	151.21	34.52
BT	0.92	171.47	108.89	24.28	0.90	195.23	125.80	27.85
XGBTree	0.96	127.63	85.21	18.07	0.91	183.80	119.59	26.22

Figura 4: Detalle de los resultados del artículo tres..

III. METODOLOGÍA Y EXPERIMENTOS

En este estudio es importante comparar los resultados obtenidos de los diferentes algoritmos convencionales de aprendizaje automático, con el fin de destacar los aspectos positivos y negativos de cada uno de los modelos que se consideraron en el presente estudio. Se consideraron cuatro

algoritmos de predicción para comparar su desempeño entre sí.

(1) Regresión múltiple: La regresión lineal supone que la relación entre dos variables tiene una forma lineal (o linealizable mediante alguna transformación de las variables). La regresión lineal tiene una versión "simple" que empareja dos variables, pero esta suele ser insuficiente para entender fenómenos mínimamente complejos en la que influyen más de dos variables, esta versión es la "múltiple". En el modelo de regresión lineal múltiple suponemos que más de una variable tiene influencia o está correlacionada con el valor de una tercera variable[5].

(2) Random Forest: Los modelos Random Forest están formados por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo[6].

(3) Redes Neuronales Artificiales: La red neuronal más utilizada se denomina Perceptrón Multicapa o MLP (Multi-Layer Perceptrón). Esta es una red de varias capas, usualmente tres (entrada, oculta y salida) que utiliza funciones sigmoideas como función de transferencia en la capa oculta. Las funciones de la capa de salida pueden ser lineales o sigmoidales, dependiendo del tipo de salida que se quiera. Pero la característica más importante de la MLP es que utiliza como función de aprendizaje la Retropropagación o regla Back Propagation (BP)[7].

(4) Regresión por Vectores de Soporte con kernel RBF: El modelo desarrollado en la década de los 90, dentro del campo de la ciencia computacional. Si bien originariamente se desarrolló como un método de clasificación binaria, su aplicación se ha extendido a problemas de clasificación múltiple y regresión. *SVMs* ha resultado ser uno de los mejores clasificadores para un amplio abanico de situaciones, por lo que se considera uno de los referentes dentro del ámbito de aprendizaje estadístico y machine learning[8].

Con los anteriores modelos de aprendizaje automático se entrenaron dos modelos uno que corresponde a Holidays y otro que corresponde a los Weekdays esto debido a que en los días festivos la variación en el servicio es más alta. Con respecto a la validación se usó KFolds con 10 splits por medio del grid search y como medidas de desempeño se evaluaron el R^2 ,RMSE(Error cuadrático medio) y el MAE(Error Absoluto Medio). De las cuales se utilizó el R^2 como medida principal de comparación con los artículos investigados. Los resultados de los experimentos se muestran en la Tabla 3 y 4.

	Holidays					
	Training			**Testing**		
	R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**
(1)	0.68	234.2	310.37	0.62	276.29	385.58
(2)	0.96	66.4	98.44	0.89	136.93	202.86
(3)	0,80	161.50	246.98	0.70	230.27	342.36
(4)	-0.23	393.03	614.28	-0.22	433.54	698.84

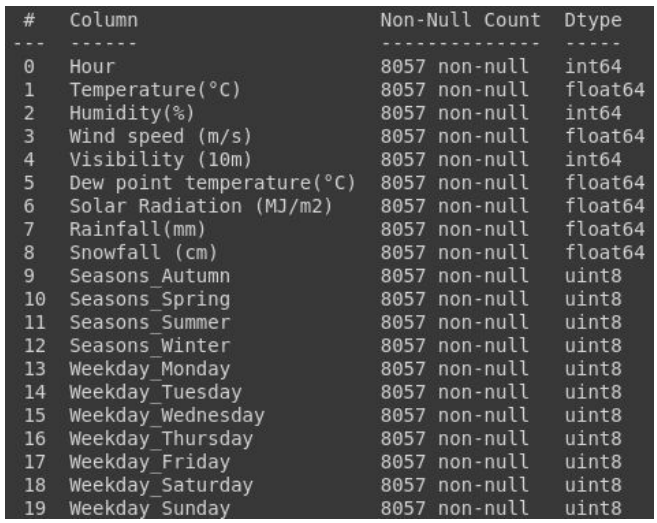
Tabla 3: Resultados de los experimentos para el modelo Holidays.

	Weekdays					
	Training			**Testing**		
	R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**
(1)	0.54	324.43	433.67	0.52	333.16	442.93
(2)	0.83	170.89	259.02	0.79	188.10	287.87
(3)	0.68	247.38	364.08	0.67	248.71	367.13
(4)	-0.06	502.59	665.87	-0.05	492.38	661.09

Tabla 4: Resultados de los experimentos para el modelo Weekdays.

```
#   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
0   Hour                       408 non-null    int64
1   Temperature(°C)            408 non-null    float64
2   Humidity(%)                408 non-null    int64
3   Wind speed (m/s)           408 non-null    float64
4   Visibility (10m)           408 non-null    int64
5   Dew point temperature(°C)  408 non-null    float64
6   Solar Radiation (MJ/m2)    408 non-null    float64
7   Rainfall(mm)               408 non-null    float64
8   Snowfall (cm)              408 non-null    float64
9   Seasons_Autumn             408 non-null    uint8
10  Seasons_Spring             408 non-null    uint8
11  Seasons_Summer             408 non-null    uint8
12  Seasons_Winter             408 non-null    uint8
13  Weekday_Monday             408 non-null    uint8
14  Weekday_Tuesday            408 non-null    uint8
15  Weekday_Wednesday          408 non-null    uint8
16  Weekday_Thursday           408 non-null    uint8
17  Weekday_Friday             408 non-null    uint8
18  Weekday_Saturday           408 non-null    uint8
19  Weekday_Sunday             408 non-null    uint8
```

Figura 5: Importancia de variables para el modelo holiday.



#	Column	Non-Null Count	Dtype
0	Hour	8057 non-null	int64
1	Temperature(°C)	8057 non-null	float64
2	Humidity(%)	8057 non-null	int64
3	Wind speed (m/s)	8057 non-null	float64
4	Visibility (10m)	8057 non-null	int64
5	Dew point temperature(°C)	8057 non-null	float64
6	Solar Radiation (MJ/m2)	8057 non-null	float64
7	Rainfall(mm)	8057 non-null	float64
8	Snowfall (cm)	8057 non-null	float64
9	Seasons_Autumn	8057 non-null	uint8
10	Seasons_Spring	8057 non-null	uint8
11	Seasons_Summer	8057 non-null	uint8
12	Seasons_Winter	8057 non-null	uint8
13	Weekday_Monday	8057 non-null	uint8
14	Weekday_Tuesday	8057 non-null	uint8
15	Weekday_Wednesday	8057 non-null	uint8
16	Weekday_Thursday	8057 non-null	uint8
17	Weekday_Friday	8057 non-null	uint8
18	Weekday_Saturday	8057 non-null	uint8
19	Weekday_Sunday	8057 non-null	uint8

Figura 6: Importancia de variables para el modelo weekday.

IV DISCUSIÓN

Las tablas 3 y 4 muestran el rendimiento de los modelos desarrollados tanto en el conjunto de datos de entrenamiento como en el conjunto de datos de prueba. El modelo que proporciona los valores R^2 más altos y el RMSE y MAE más bajos es el mejor.

Como se puede observar, Random Forest y el modelo basado en redes neuronales artificiales muestran mejores resultados en el entrenamiento, pero el rendimiento del Random Forest fue ligeramente superior al MLP en el conjunto de pruebas. Dado que el conjunto de pruebas se considera como el criterio de rendimiento final para cualquier modelo de regresión, en este caso el modelo RF tiene el mejor rendimiento que otros modelos de regresión. Además se puede observar que el peor modelo es el basado en SVM, este modelo en especifico es peor que predecir la media.

V.CONCLUSIONES

Utilizar el RF en la predicción de la demanda de bicicletas de alquiler y comparar el rendimiento de la predicción con los métodos tradicionales, es decir, SVM y RNN.

El pronóstico de demanda de alquiler de bicicletas de este artículo, el modelo de regresión múltiple y SVM convencional no es aplicable. Según este informe se propone un modelo de demanda de alquiler de bicicletas basado en bosque aleatorio.

Lo cual se ve también reflejado en los artículos de soporte en los cuales una metodología basada en árboles de decisión se encontraba entre los mejores métodos.

REFERENCIAS

[1] VE, S., & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 1-18.

[2] VE, S., & Cho, Y. Season wise bike sharing demand analysis using random forest algorithm. *Computational Intelligence*.

[3] Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, 353-366.

[4] Feng, Y., & Wang, S. (2017, May). A forecast for bicycle rental demand based on random forests and multiple linear regression. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)* (pp. 101-105). IEEE.

[5] Granados, R. M. (2016). Modelos de regresión lineal múltiple. *Granada, España: Departamento de Economía Aplicada, Universidad de Granada*.

[6] Árboles de decisión, random forest, gradient boosting y C5.0 by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_C50.html

[7] Pérez Ramírez, F. O., & Fernández Castaño, H. (2007). Las redes neuronales y la evaluación del riesgo de crédito. *Revista Ingenierías Universidad de Medellín*, *6*(10), 77-91.

[8] Máquinas de Vector Soporte (Support Vector Machines, SVMs) by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines