MEMORIA JUSTIFICATIVA PARA LA ACEPTACIÓN DEL ANTEPROYECTO

Título: Modelo predictivo de la demanda de viajeros del Metro de Madrid

Autor: Carlos de los Ríos Mouvet

DESCRIPCIÓN DEL PROBLEMA A RESOLVER

Motivación y origen

La red de Metro de Madrid es una de las grandes infraestructuras tanto de la ciudad de Madrid como de España e incluso Europa, siendo la tercera red más grande del continente en cuanto a longitud de vía. Este medio de transporte que irrumpió en la vida de los madrileños a principios del siglo XX se ha convertido en un elemento vital de la ciudad, es por ello por lo que se justifica todo esfuerzo dedicado a mantener y mejorar una infraestructura tan importante como Metro de Madrid.

No es de extrañar que una red de semejante tamaño requiera unos niveles de inversión muy considerables para mantener un funcionamiento eficiente y los proyectos de ampliación que se han ido sucediendo desde que en 1919 se inaugurasen las 8 estaciones originales. En este aspecto también hay que entender que para que la inversión llegue a buen puerto debe cumplir con su objetivo principal: llevar al mayor número de viajeros posible. Es por ello por lo que también es interesante conocer cómo afectan diferentes características de un barrio a la hora de que sus vecinos usen el Metro, de manera que antes de empezar a excavar un solo metro cúbico de tierra ya podamos estimar cuánta gente usará esa parada.

En términos de seguridad ciudadana, la Ley 8/2011, de 28 de abril, por la que se establecen medidas para la protección de las infraestructuras críticas define las mismas de la siguiente forma:

> *"Infraestructuras críticas: las infraestructuras estratégicas cuyo funcionamiento es indispensable y no permite soluciones alternativas, por lo que su perturbación o destrucción tendría un grave impacto sobre los servicios esenciales."*

A pesar de que, por motivos evidentes, no existe una lista oficial de las infraestructuras críticas de España, tenemos la certeza de que la red de Metro de Madrid forma parte de la misma gracias a la respuesta a una solicitud de información firmada por el responsable del área de cumplimiento normativo y transparencia en 2017. Este nivel de protección no es casual ya que, además de la relevancia en materia social y económica que he expuesto antes, Metro de Madrid ha sido objetivo tanto de amenazas terroristas (afortunadamente no ejecutadas) como de sabotajes. Además, debemos tener en cuenta que no solo una acción premeditada puede poner en peligro el normal funcionamiento de la red de Metro, su propio "éxito" puede ser causa de colapsos parciales o totales de la misma, por lo que es preciso conocer la demanda a la que se enfrentará en futuro para analizar la capacidad de respuesta.

De la importancia para la logística, la economía y la seguridad ciudadana (entre otros muchos ámbitos) de la Comunidad de Madrid se desprende la motivación de este trabajo, en el que se pretende obtener una predicción fiable y precisa de la demanda de viajeros de la red de Metro de Madrid.

Datos que lo sustentan

Durante el pasado año 2023 se realizaron 662.273.698 viajes por los casi 300 kilómetros y las 302 estaciones del Metro de Madrid. Esto supone en torno al 42% de los viajes realizados en el sistema de transporte público de la Comunidad de Madrid. En términos económicos, la Comunidad Autónoma de Madrid, accionista única de Metro, tiene previsto destinar el 5,177% (1.484 millones de euros) de sus Presupuestos Generales 2025 a los ferrocarriles metropolitanos. Analizando la cuenta de Pérdidas y Ganancias de Metro de Madrid S.A. observamos que los ingresos derivados del número de viajeros en el ejercicio 2023 ascienden a 1.011,657 millones de euros. De esos resultados, 992,86 millones provienen de manera directa del transporte de viajeros (El Consorcio Regional de Transportes de Madrid abona una tarifa a Metro por cada viaje de 1,65€, el resto lo abona el usuario) mientras que otros 24,55 millones provienen de la comisión por venta títulos del CRTM en la red de Metro.

Impacto de la solución del problema

Los sistemas de gestión de transporte han resultado ser en los últimos años herramientas vitales para los grandes operadores de transportes. Como hemos visto anteriormente, la red de Metro de Madrid precisa de unos niveles de sofisticación muy elevados debido a su relevancia social y económica.

Mediante una herramienta de predicción de demanda el operador puede disponer de información clave a la hora de planificar los recursos humanos, materiales y económicos necesarios para el perfecto funcionamiento.

Esta herramienta también tiene impacto en otra fuente de ingresos de Metro: la publicidad. De esta forma Metro de Madrid puede personalizar aún más sus servicios publicitarios de manera que un anunciante pueda saber con exactitud donde va a tener más impacto su campaña.

Problema de investigación en forma de Pregunta

¿Cómo podemos predecir de manera precisa la demanda de viajeros de Metro de Madrid?

Preguntas de investigación.

¿Qué características tiene cada estación de la red de Metro de Madrid?

¿Qué variables explican mejor la demanda de viajeros de Metro de Madrid?

¿Qué modelo predice mejor la demanda de viajeros de Metro de Madrid?

Estado del arte

WALTERS Y CERVERO (2003), propusieron en su trabajo un modelo directo de estimación de la demanda a nivel de estación. En este caso se trata de una modelo de regresión múltiple, que considera de forma detallada el entorno de las estaciones.

Estas múltiples variables independientes se clasifican en externas (sociodemográficas, económicas, modos complementarios o competidores, disponibilidad de aparcamiento en la estación, espaciales, etc.) e internas (oferta de servicio, tarifas, frecuencia de servicio, puntualidad, etc.) Como norma general, las variables externas tienden a explicar mejor que las internas y, dentro de las internas, el nivel de servicio es más relevante que las tarifas (TAYLOR Y FINK, 2003)

Estudios llevados a cabo en TODs (Transit Oriented Development) situados alrededor de las estaciones de ferrocarril de California llevan a concluir que la densidad es el principal predictor del uso del ferrocarril. Al existir altas densidades, la transitabilidad peatonal y la mezcla de usos adquieren una importancia menor (CERVERO, 2004)

En estas conclusiones se basa el trabajo Gutiérrez Puebla, J., Cardozo, O. D., & García Palomares, J. C. (2008). Modelos de demanda potencial de viajeros en redes de transporte público: aplicaciones en el Metro de Madrid. En este trabajo se propone una metodología basada en Sistemas de Información Geográfica para estimar la demanda de viajes en estaciones de la red de metro de Madrid.

La particularidad que aporta este trabajo es uso de un modelo de curva de demanda basado en la distancia a pie de una estación como un factor de ponderación de las variables del Modelos de Regresión Múltiple que hacen referencia al entorno de las estaciones. Al realizar los análisis se obtuvieron mayores correlaciones con una agrupación de la información en un radio de 500 metros. Es a partir de esa distancia cuando las correlaciones empiezan a reducirse, excepto en los casos de estaciones céntricas con poca distancia unas de otras donde el radio es menor.

Las relaciones más elevadas en la matriz de correlación de Pearson (coeficientes por encima de 0.5) se dan en variables económicas de la población, como el empleo en educación, sanidad y en comercio, y la distribución de dicha población, ya sea total u ocupados. También existe una relación clara entre la propia población total y los ocupados y la población total y el número de estudiantes y, de una manera significativa, una relación entre número de estudiantes y el empleo en educación. Ya en las variables utilizadas para la definir a la población se encuentra correlación entre el porcentaje de extranjeros y la proporción de hogares sin coche.

Si entramos en el aspecto de las variables de diseño territorial, la densidad viaria tiene una correlación alta con la distribución de la población, e igualmente aparece correlación negativa entre la distancia al centro y los hogares sin coche o la distancia al centro y la densidad viaria.

A raíz de estos resultados se propone un primer modelo en el que se incorporan las variables: población total, empleo total, densidad de vivienda multifamiliar, densidad viaria, diversidad de usos, número de líneas y número de bocas de la estación. Los coeficientes β estandarizados muestran que las variables que mejor explican son aquellas relacionadas con el nivel de servicio: número de líneas y número de bocas (en este caso a través de áreas de cobertura mayores). Con menor poder explicativo se sitúan las variables de empleo y población total.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,888(a)	,789	,776	67728,991
2	,888(b)	,789	,777	67499,833
3	,887(c)	,787	,777	67489,182
4	,886(d)	,785	,776	67643,921
5	,885(e)	,783	,776	67733,509

La ecuación del modelo final seleccionado, al que se le realizarían los testados de errores es la siguiente:

$$ENTRADAS = -82477{,}290 + (0{,}563 * EMPTOT) + (1{,}059 * POBTOT) + (1745{,}988 * POBEXT) + (24518{,}508 * NUM_BOCA) + (150854{,}88 * NUM_LIN)$$

Este modelo presenta un error medio según estación del ±24.8%, con una cierta tendencia a la sobreestimación de las entradas, sin que se aprecien patrones en los errores Los datos son los siguientes:

Tabla 16: Estimación de los errores según línea.

Línea	Estaciones consideradas	Entradas observadas	Entradas estimadas	Diferencia	Diferencias relativas
1	22	6702243	6241817	-460426	-6.87
2	9	1931514	2483514	552000	28.58
3	7	3026240	2421845	-604395	-19.97
4	15	4086851	4547214	460363	11.26
5	14	3198750	3169617	-29133	-0.91
6	15	5075123	4460182	-614941	-12.12
7	18	3835249	4083068	247819	6.46
8	4	1144196	1138016	-6180	-0.54
9	19	3367346	3697664	330318	9.81
10	8	2059344	1591411	-467933	-22.72
11	2	167860	265008	97148	57.87
12	21	2280063	2794630	514567	22.57

Basado también en este trabajo encontramos el de (Villalba Pérez, 2018), cuya particularidad respecto al anterior es la utilización de árboles regresivos para obtener un modelo predictivo ajustado para realizar estimaciones de nuevas estaciones extramuestrales.

En cuanto a la consideración de variables se tienen en cuenta algunas que Cardozo et al. (2008) no incluyen en el modelo, tales como: Renta per Cápita y Renta media por hogar, Transbordo con Red de Transporte Externa, Estación Accesible, Estación con Parking, Zona de Influencia Turístico- Cultural-Comercial y Número de Estaciones Restantes hasta Transbordo.
En este trabajo también se plantea un análisis en las distribuciones de los usuarios de líneas del metro para estimar si la línea "admite" una nueva parada. Para ello se considera que, si la distribución de viajeros dentro de la línea sigue una distribución normal, la línea no aceptaría una nueva parada, puesto que los viajeros que la usen serían residuales. Una hipótesis que, a mi criterio, no se sostiene ya que considera que el único factor relevante para el uso de una estación es su cercanía al centro.

Para el análisis de la regresión se plantean tres conjuntos diferenciados: **Conjunto 1**, total de estaciones excluyendo outliers (Ciudad Universitaria, Aeropuerto T1-T2-T3, Aeropuerto T4 y Feria de Madrid); **Conjunto 2**, estaciones de la periferia (situadas fuera de la almendra central/M-30); y **Conjunto 3**, estaciones en los extremos de líneas. Los resultados para cada conjunto son los siguientes:

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	0.310	0.091	3.408	0.001
I2.RentaPC	0.065	0.029	2.265	0.025
I3.Transbordo_metro	0.135	0.020	6.862	1.05e-10
I4.Transbordo_Externo	0.151	0.019	7.956	1.87e-13
I5.Estacion_adaptada	0.024	0.015	1.552	0.123
I14.DensidadPobl_Hogar	-0.107	0.041	-2.613	0.010
I19.Tasa_Inmigracion	-0.004	0.002	-2.307	0.022

Residual standard error	0.093 on 181 degrees of freedom
Multiple R-squared	0.567
Adjusted R-squared	0.552
F-statistic	39.45 on 6 and 181 Degrees of freedom, p-value: < 2.2e-16

[Figura 9] Regresión ajustada del Conjunto 1

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	-0.395	0.296	-1.336	0.185
I3.Transbordo_metro	0.284	0.054	5.200	1.26e-06
I4.Transbordo_Externo	0.171	0.037	4.636	1.21e-05
I5.Estacion_adaptada	-0.043	0.027	-1.582	0.117
I6.Estacion_parking	0.110	0.049	2.263	0.026
I8.Poblacion_Joven	0.338	0.066	5.088	2.00e-06
I12.Poblacion_edad_trabajar	0.702	0.363	1.935	0.056
I13.Zona_Envejecida	0.059	0.040	1.471	0.145
I15.Estaciones_Hasta_Transbordo	-0.011	0.006	-1.831	0.070
I19.Tasa_Inmigracion	-0.004	0.002	-1.878	0.063

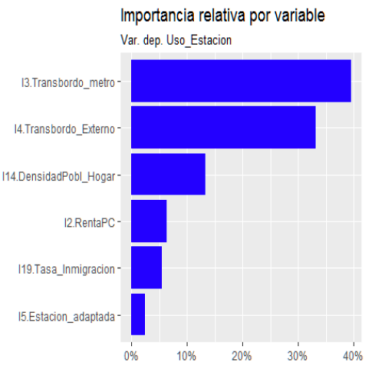
Residual standard error	0.115 on 89 degrees of freedom
Multiple R-squared	0.677
Adjusted R-squared	0.644
F-statistic	20.74 on 9 and 89 Degrees of freedom, p-value: < 2.2e-16

[Figura 12] Regresión ajustada Conjunto 2

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	-0.064	0.054	-1.196	0.235
I2.RentaPC	0.121	0.046	2.652	0.009
I3.Transbordo_metro	0.150	0.030	5.057	1.80e-06
I4.Transbordo_Externo	0.151	0.025	5.966	3.25e-08
I5.Estacion_adaptada	0.035	0.022	1.597	0.113
I7.Influencia_Comercial	0.054	0.030	1.788	0.077
I8.Poblacion_Joven	0.087	0.052	1.691	0.094
I13.Zona_Envejecida	0.037	0.021	1.730	0.090
I19.Tasa_Inmigracion	-0.004	0.002	-2.253	0.026

Residual standard error	0.100 on 106 degrees of freedom
Multiple R-squared	0.634
Adjusted R-squared	0.607
F-statistic	22.99 on 8 and 106 Degrees of freedom, p-value: < 2.2e-16

[Figura 16] Regresión del Conjunto 3



Analizando los resultados por conjunto podemos concluir que en el Conjunto 1 se obtienen resultados similares a Cardozo et al. (2008) en las que la densidad poblacional, la renta per cápita (la cual guarda estrecha relación con el nivel de empleo) y la tasa de inmigración influyen en la demanda. Sin embargo, la accesibilidad de la estación no parece ser significativa.

Lo interesante de la aproximación realizada por Villalba Pérez (2018) es la evolución de los factores en función de las características. Como vemos en el conjunto 2 (los más alejados de la almendra central) la densidad poblacional pierde mucha relevancia en favor de otras variables como el nivel de población joven y la presencia de aparcamiento (intermodalidad). En este conjunto las variables población en edad de trabajar, estaciones hasta transbordo o tasa de inmigración no parecen ser significativas.

A pesar de que el trabajo de Cardozo et al. (2008) resulta ser más robusto y eficaz a la hora de explicar la demanda podemos sacar una conclusión clara: la concurrencia de varias líneas y/o métodos de transporte en una estación es la principal explicación de la demanda en una estación de la red de Metro de Madrid.

En cuanto a literatura acerca de estudios realizados en otros sistemas de transporte en otros países es interesante tomar en cuenta el trabajo de Ling, X., Huang, Z., Wang, C., Zhang, F., & Wang, P. (2018). Predicting subway passenger flows under different traffic conditions. En este estudio el equipo chino utiliza los datos de la red de metro de Shenzhen para poner a prueba cuatro modelos: modelo de media histórica (HA), perceptrón multicapa (MLP), Vectores de Soporte Regresión (SVR) y árboles de regresión con gradient boosting (GBRT)
A diferencia de otros sistemas como el madrileño, los usuarios del Metro de Shenzhen han de validar su título también a la salida, este hecho facilita bastante el análisis de los datos de los viajeros.
Los resultados revelan que, en condiciones normales, el modelo SVR es capaza de lograr el mejor RMSE con 32,57 respecto al 34,27 del modelo MLP y al 37,34 del GBRT. Es cuando nos encontramos en circunstancias anómalas (caídas o picos de demanda atípicos) donde mejor trabaja el modelo MLP donde consigue obtener un RMSE de 58,17 frente al 65,71 del modelo SVR.

En el estudio Halyal, S., Mulangi, R. H., & Harsha, M. M. (2022). Forecasting public transit passenger demand: With neural networks using APC data, el equipo investigador se enfoca en predecir la demanda de viajeros del sistema de tránsito rápido de autobuses Hubballi-Dharwad, en India, utilizando el modelo de aprendizaje automático Long Short-Term Memory (LSTM) En este estudio también se usó el modelo Seasonal Autorregressive Integrated Moving Average (SARIMA) y el modelo Seasonal Naive para comparar resultados a fin de entender que modelo ajustaba mejor la realidad.

Para este estudio se recogieron 54.270.301 observaciones de entradas únicas a toda la red entre el 01/12/2019 y el 29/02/2020. Para el entrenamiento de los modelos se seleccionaron aquellas procedentes de las 5 estaciones con más tránsito y almacenando únicamente valores de fecha, hora y número de usuarios en tramos de media hora desde las 06:00. Tras la construcción de los modelos se calculó para los tres sus Mean Absolute Error (MAE) y Root Mean Square Error (RMSE) a fin de comparar la precisión de la triada de modelos, los resultados fueron los siguientes:

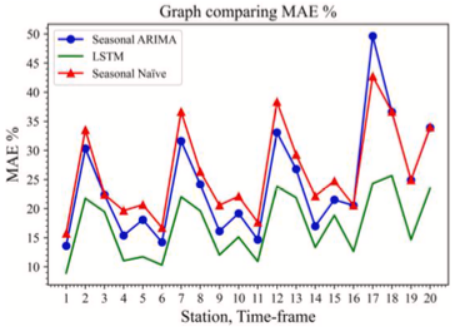


Fig. 3. Graph comparing MAE %.

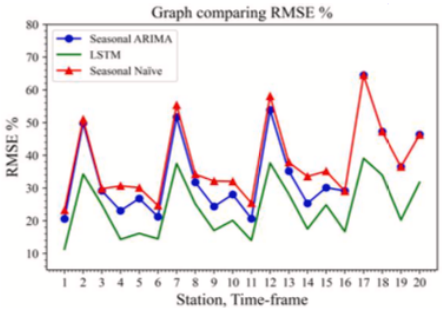


Fig. 4. Graph comparing RMSE %.

Los resultados de la comparación no dejan lugar a duda: el modelo LSTM es capaz de explicar mejor la demanda de pasajeros del sistema de tránsito rápido de autobuses Hubballi-Dharwad.

Marco teórico

Regresión Lineal Múltiple

El análisis de la regresión es un proceso estadístico en el que tratamos de explicar como una variable depende de otra variable. Este proceso fue desarrollado por primera vez por Legendre (1805) y por Gauss (1809) a través del método de mínimos cuadrados con el objetivo de determinar, a partir de observaciones astronómicas, las órbitas de los cuerpos alrededor del Sol.

Dentro de los diferentes modelos de análisis de regresión nos encontramos con la Regresión Lineal Múltiple (RLM) el cual se enfoca a describir la relación de dependencia entre una variable dependiente Y, k variables independientes X y un término aleatorio. Este modelo se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Y: variable que tratamos de predecir o explicar

$X_1, X_2, \dots, X_k$: variables que explican el valor de Y

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$: parámetros del modelo que definen la importancia de las variables X

ε: error medio

Cuando operamos este modelo es habitual expresarlo de manera matricial para un manejo más cómodo de la siguiente manera:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \underset{n\times(p+1)}{\mathbf{X}} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \underset{(p+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \underset{n\times 1}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

por ello podemos expresar la función de regresión lineal múltiple de la siguiente forma: $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}} \underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{\epsilon}$

Para obtener los parámetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ debemos utilizar el método de mínimos cuadrados:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Para aplicar correctamente este modelo debemos cumplir una serie de hipótesis sobre ε:

- $E[\varepsilon_i] = 0$: el objetivo es que el error, de media, sea 0
- $V[\varepsilon_i] = \sigma^2$: homocedasticidad, la dispersión que presentan los datos debe permanecer constante para todos los valores que puede tomar X
- $Cov(\varepsilon i, \varepsilon j) = 0,\ \forall i \neq j$. lo que haya sucedido en la observación i-ésima no influye en otra observación j-ésima.

Aunque estas hipótesis sean requisitos que ha de cumplir el error medio, existen otras hipótesis que debe cumplir el modelo para que podamos confiar en las predicciones o interpretaciones de este:

- **Validez**: Los valores de la variable objetivo deben medir de forma precisa el fenómeno a estudiar. El modelo debe incluir todos los predictores relevantes
- **Representatividad**: Los datos de la muestra deben ser representativos de la población en la que se vaya a utilizar el modelo
- **Aditividad y Linealidad**: la relación entre los predictores y la variable objetivo es lineal, si no se da este caso la capacidad predictiva disminuirá y la interpretación será errónea.
- **Normalidad de los errores**: el conjunto de errores ε_i ha de seguir una distribución normal.

Test Shapiro-Wilk

Como hemos visto anteriormente, uno de los requisitos del modelo de regresión es la normalidad de los errores, esto es: que el conjunto de los errores registrados en cada observación siga una distribución normal. La prueba de Shapiro-Wilk es una prueba estadística que se utiliza para determinar la normalidad de un conjunto. Para ello se utiliza un estadístico W encargado de evaluar la correlación entre los datos y los valores esperados en una distribución normal. Un valor W cercano a 1 sugiere que los datos son normales. La fórmula es la siguiente:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Coeficiente de correlación de Pearson

Indica el grado de relación lineal entre dos variables cuantitativas. Toma valores entre [-1, +1], siendo +1 una correlación lineal positiva perfecta y -1 una correlación lineal negativa perfecta. Si es 0 significa que NO existe relación lineal entre las variables consideradas. La fórmula es la siguiente:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Long Short Term Memory

Dentro de los diferentes modelos que podemos encontrar en el espectro del aprendizaje automático (Machine Learning) nos encontramos con las redes neuronales artificiales, las cuales tratan de replicar la estructura y funciones de las neuronas biológicas de los cerebros animales.

Los inicios de este campo los desarrollaron McCulloch y Pitts en 1943 al crear un primer modelo neuronal que modelase el comportamiento de una neurona humana. Ya a finales de la década de los años 40 sería Donald Hebb el que desarrollase una hipótesis de aprendizaje apoyándose de su especialidad, la psicología. Más adelante, en 1958, Frank Rosenblatt se inspiraría en el trabajo de McCulloch y Pitts para crear el perceptrón simple, un algoritmo de identificación de patrones el cual realiza predicciones basadas en una combinación lineal de ciertos pesos y entradas. Este modelo de Red Neuronal Artificial es considerado como el pionero dentro del campo y, a pesar de ello, sigue utilizándose hoy en día con muy buenos resultados.

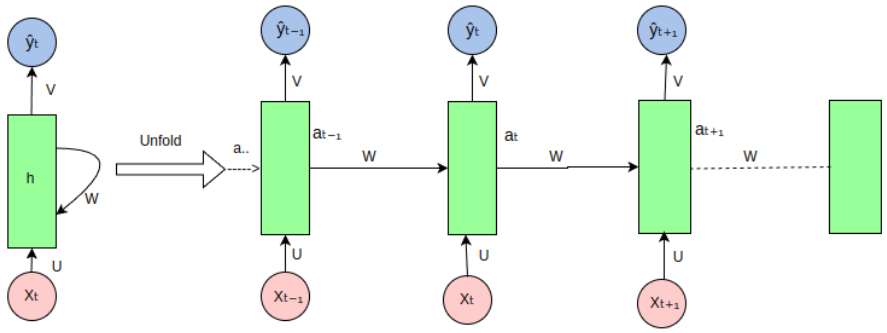
A finales de los años 70, Minsky y Papert (1969) demostraron que el perceptrón simple no podía resolver problemas no lineales, esta limitación suponía que la red neuronal fuese incapaz de aprender problemas sencillos como una función lógica XOR. Este fue el germen del trabajo de Rumelhart, Hinton y Williams (1986), los cuales desarrollaron el algoritmo de retro propagación (backpropagation) que solucionaba las limitaciones del perceptrón simple. Este concepto de perceptrón multicapa resultó dar un gran impulso a este tipo de modelos de aprendizaje automático.

Es a raíz del trabajo de Rumelhart et al. (1986) cuando surgen las Redes Neuronales Recurrentes (RNN, su acrónimo en inglés), las cuales son redes neuronales de aprendizaje automático supervisado con uno o más bucles de retroalimentación. Dentro del espectro de las RNN podemos encontrar las redes de Hopfield (Hopfield, 1982)

Dentro de las RNN, Hochreiter y Schmidhuber (1997) desarrollaron las redes Long Short-Term Memory (LSTM). Este tipo de redes destaca por tener celdas de memoria que permiten a la red recordar estados por períodos cortos o largos de manera que ayuden a decidir el siguiente estado. Esta característica es la que hace que las LSTM sean idóneas para el procesamiento de datos en series temporales.

Arquitectura de una LSTM

Como ya hemos visto anteriormente una red LSTM es una RNN con unas particularidades de carácter más bien interno. Toda RNN se compone de una serie de neuronas con la siguiente estructura:

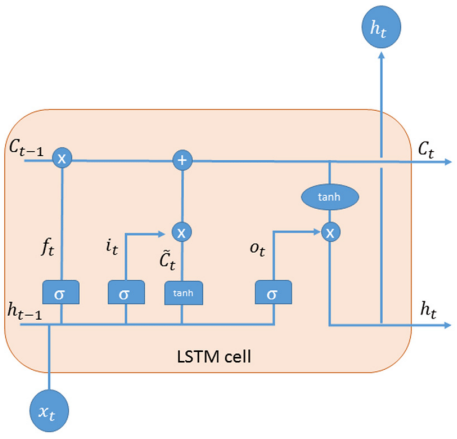


El proceso de aprendizaje comienza con la inserción de un vector x_t a la red, la red escanea el vector anterior x_{t-1} y el siguiente x_{t+1}, actualiza el estado oculto a_t y devuelve un vector y_t y el estado oculto a_t actualizado a la siguiente neurona.

Para actualizar el estado oculto la neurona toma el input x y el estado oculto anterior y realiza una transformación lineal y la función de activación tangente hiperbólica. Después genera la predicción, tomando el estado oculto ya actualizado y aplicando otra transformación que es posteriormente llevada a una función softmax.

El problema de este método reside en su función de activación ya que las funciones tangentes hiperbólicas internas se van anidando por lo que el efecto que tiene un estado oculto a partir de la tercera salida será mínimo.

De esta problemática surge la propuesta de Hochreiter y Schmidhuber (1997) de las Long Short-Term Memory



Si comparamos esta arquitectura con la de una neurona RNN vemos un nuevo elemento C, el cual hace referencia a la celda de estado. La celda de estado es el competente clave de una neurona LSTM pues es la encargada almacenar y regular la información de manera que una neurona pueda retener o eliminar información según las necesidades de esta. Para ello la neurona dispone de la puerta de olvido f_t (encargada de eliminar información no útil), la puerta de entrada i_t (encargada de añadir nuevos elementos a la memoria) y la puerta de salida o_t (encargada de crear el nuevo estado oculto)

OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS

Objetivo General

Predecir de manera precisa la demanda de viajeros de Metro de Madrid

Objetivos Específicos

Definición de características de cada estación de la red de Metro de Madrid

- Características demográficas
- Características propias de la estación (Transbordo, bocas, etc.)
- Estación climatológica asociada

Normalización de variables

Cálculo de variables que explican mejor la demanda de viajeros de Metro de Madrid

Modelo Regresión Lineal Múltiple

Modelo LSTM

Comparación Modelos

METODOLOGÍA

Proceso ETL

Extracción

Para el proceso de elaboración de datos parto de aquellos datos que formarán la variable dependiente del modelo. El proceso de obtención será a través del portal de datos abiertos de Metro de Madrid. En este portal podemos encontrar los siguientes datasets:

- Utilizaciones por Estaciones (Suma de entradas, salidas y transbordos de una estación)
- Demanda Mensual y Acumulada (Suma de entradas a todas las estaciones de la red)
- Demanda diaria (Suma de entradas a todas las estaciones de la red)
- Entradas mensuales por estación

Así mismo este portal ofrece información sobre el número de reclamaciones de toda la red según su motivo y el registro de obras en la red. También incluyen informes anuales de la demanda con diferentes insights.

Respecto a la obtención de datos climatológicos la fuente principal será la API de la Agencia Estatal de Meteorología. En esta API podemos encontrar datos históricos tanto de las predicciones de cada día como de los registros reales.

Para la obtención de los precios de los carburantes en la ciudad de Madrid se usará la herramienta desarrollada por el Ministerio para la Transición Ecológica y el Reto Demográfico, la cual facilita datos diarios de cada estación de servicio de España. A fin de no añadir complejidad a este proceso (ya complejo de por sí) se tendrá en cuenta el precio promedio diario del Gasóleo A habitual (Diesel) y de la Gasolina 95 E5 en las estaciones de servicio de la provincia de Madrid.

Otras fuentes de datos serán:

- CRTM
- Google Earth
- INE
- Ayuntamiento de Madrid
- Comunidad de Madrid

Podrán encontrar un anexo con diferentes muestras de los datos mencionados.

Transformación

En este apartado se procederá a “limpieza” del dato. Es en este momento en el que se tratarán los valores nulos, las entradas repetidas y la transformación de variables cualitativas si el modelo lo requiere.

Carga

Con los datos ya tratados correctamente se procederá a la carga de los mismo en el sistema de bases de datos óptimo para el conjunto.

Análisis Exploratorio

Una vez realizado el proceso ETL el siguiente paso consiste en analizar y comprender la naturaleza de los datos con los que contamos. Es en esta fase donde se van a realizar las identificaciones de patrones estacionales, tendencias significativas, valores atípicos, etc.

En este proyecto se pretende comparar el funcionamiento de dos modelos predictivos los cuales requieren que los datos para su entreno tengan características concretas. Es por ello por lo que en este punto también consiste en realizar las diferentes pruebas para cerciorase de que los datos son válidos para los modelos elegidos.

Construcción de modelos

Con los datos ya revisados se procederá a la construcción de los dos modelos propuestos: Regresión Lineal Múltiple y Long Short-Term Memory (LSTM) En este apartado los esfuerzos se centran exclusivamente en armar un modelo descriptivo de los datos históricos.

Evaluación y validación de resultados

Una vez construidos ambos modelos se procederá verificar los resultados obtenidos por ambos modelos y, si ambos obtienen resultados positivos, se compararán para evaluar cuál de los dos modelos ha resultado ser preciso.

Predicción de datos futuros e impacto

Con el modelo elegido se procederá a realizar simulaciones para obtener escenarios futuros. Una vez que se hayan obtenido estas predicciones se realizará el estudio de consecuencias

Memoria y Presentación

Una vez finalizado el trabajo "técnico" se procederá a la redacción de la memoria en la que se recogerán aprendizajes y resultados del proyecto. A su vez se trabajará en la presentación para la defensa ante el tribunal.

CRONOGRAMA

Primera fase - Proceso ETL
Duración estimada: 3 semanas
Entregables: bases de datos listas para entrenar modelos.

Segunda fase - Análisis Exploratorio
Duración estimada: 2 semanas
Entregables: informe de visualización de los datos.

Tercera fase - Construcción de modelos
Duración estimada: 7 semanas
Entregables: Modelos de Regresión Lineal Múltiple y Long Short-Term Memory (LSTM) listos para ser evaluados.

Cuarta fase - Evaluación y validación de resultados
Duración estimada: 1 semana
Entregables: informe justificativo de elección de modelo.

Quinta fase - Predicción de datos futuros e impacto
Duración estimada: 2 semanas
Entregables: informe de visualización de datos.

Sexta fase - Memoria y Presentación
Duración estimada: 5 semanas
Entregables: memoria y presentación.

BIBLIOGRAFIA

Alcaide Fernández, N. (2024). *Modelo de simulación de líneas 5 y 6 del Metro de Madrid*.

Gallo, M., De Luca, G., D'Acierno, L., & Botte, M. (2019). Artificial Neural Networks for Forecasting Passenger Flows on Metro Lines.

Gutiérrez Puebla, J., Cardozo, O., Carlos, J., & Palomares, G. (2008). *Models of potential demand of passengers on public transport networks: applications in the Metro de Madrid*

Halyal, S., Mulangi, R. H., & Harsha, M. M. (2022). Forecasting public transit passenger demand: With neural networks using APC data. *Case Studies on Transport Policy*.

Ling, X., Huang, Z., Wang, C., Zhang, F., & Wang, P. (2018). Predicting subway passenger flows under different traffic conditions.

Villalba Pérez, D. (2018). *ANÁLISIS DESCRIPTIVO Y PREDICTIVO DEL USO DE LA RED DE METRO DE MADRID*.

Wang, K., Wang, P., Huang, Z., Ling, X., Zhang, F., & Chen, A. (2022). A Two-Step Model for Predicting Travel Demand in Expanding Subways.

Wei, Y., & Chen, M.-C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks.