Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapa 4. Reporte de modelación

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

04 de noviembre de 2022

Etapa 4: “Modeling”

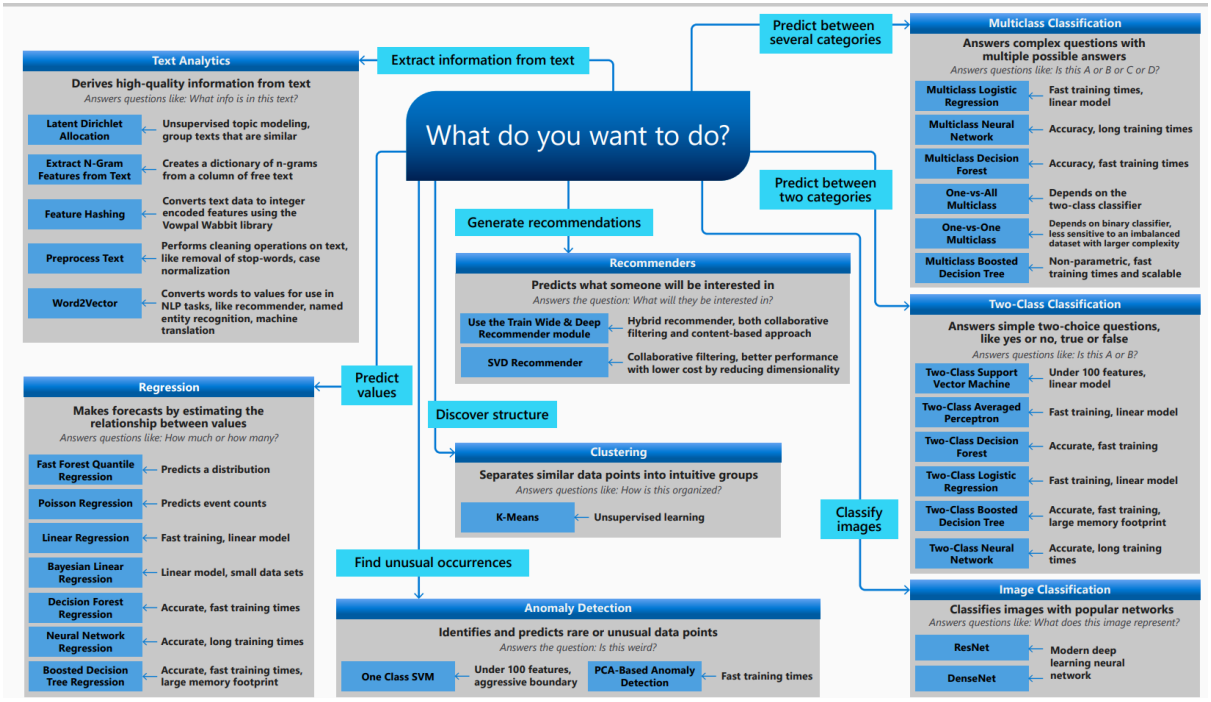
Para ejecutar esta etapa es esencial enfocarnos en el objetivo de la minería de datos, el cual es el siguiente:

“Obtener datos representativos sobre las rutas Origen-Destino mediante la limpieza, transformación y análisis del set de datos inicial. Utilizar dichos datos para entrenar un modelo que sea capaz de predecir la cantidad de viajes a una comuna en base a las características de dicha comuna.“

Por ende, resulta natural que una vez obtenida la matriz de viajes procedamos a buscar cuál es el mejor modelo para dicha predicción.

En el área de ML existen 3 grandes ramas con un gran abanico de modelos que podemos utilizar. Por lo tanto, es importante saber distinguir qué es lo que queremos hacer, para poder elegir el modelo adecuado.

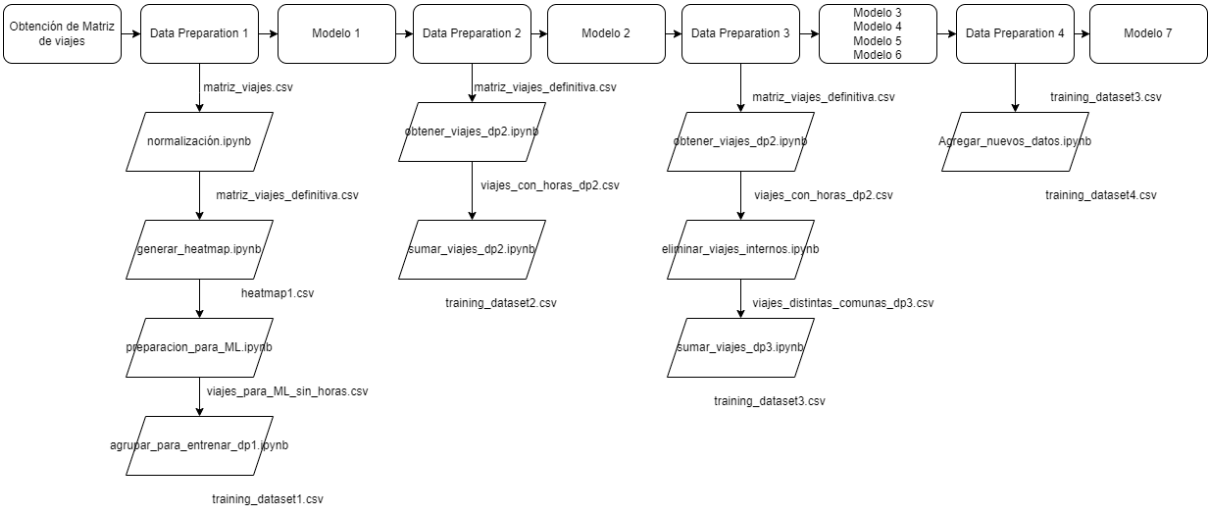
Como base para esta decisión, usaremos la guía de ML de Microsoft Azure, la cual se muestra a continuación:



En nuestro caso, queremos predecir valores (cantidad de visitas a una comuna) en base a los atributos de la comuna, por lo que necesitamos usar una Regresión.

Vemos que existen distintos tipos de regresión, cada tipo tiene distintas ventajas y requisitos para su uso.

Diagrama de modelación



Diseño de pruebas

Debido a la cantidad de registros se decidió hacer una distribución del dataset en:

- Entrenamiento: 85%
- Pruebas: 15%

Para verificar la calidad y validez de los datos se tomará como medida el porcentaje medio de error absoluto de cada dataset (MAPE).

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Se decidió optar por el uso MAPE porque se utiliza a menudo en la práctica debido a su interpretación muy intuitiva en términos de error relativo, permitiendo mostrar resultados que todos los stakeholders puedan comprender sin necesidad de ver todo el análisis previo [3].
Una de sus características es que sus distribuciones de error subyacentes de estas medidas sólo tienen valores positivos y no tienen límite superior, los errores porcentuales son muy propensos a la asimetría a la derecha en la práctica real, pero esto no debe de ser algo de que preocuparnos, ya que por el contexto no es una desventaja [1].
Una de sus principales ventajas es que aunque haya un cambio en la escala de los datos, la escala de la metrica sigue siendo la misma, siendo esto muy útil al momento de hacer distintas iteraciones en los modelos.

El MAPE ofrece las mismas propiedades que el MSE y el RMSE, pero se expresa en porcentajes [2],
Otro factor para no elegir otra métrica es que se tiene planteado el uso de variables indicatriz (dummie), este tipo de variables provocan que el uso de otras métricas como R2 se vean afectadas, provocando que su valor no sea correcto.

Para cada uno de los modelos se usarán las mismas pruebas.

Modelo 1. Regresión Lineal

Como el modelo más sencillo de regresión es el de regresión lineal, lo utilizaremos como modelo base/benchmark, y este será nuestro punto de comparación con el resto de modelos.

Además, el modelo de regresión lineal nos brinda la ventaja de que podemos analizar las propiedades estadísticas para obtener insights como la significancia de cada variable independiente.

Framework: Sci-Kit Learn

[Código Fuente]

Descripción de datos:
Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas

Y como variable dependiente:

- Número de viajes hacia la comuna

Parámetros:
Al ser un modelo de regresión lineal y se utilizará como modelo benchmark, este utiliza los parámetros por defecto de la librería.

Descripción del modelo:
Se esperan resultados no muy significativos, se estima una variación promedio mayor al 50%.

No se puede modelar relaciones complejas y no se pueden capturar relaciones no lineales sin transformar la entrada, por lo que se tiene que trabajar duro para que se ajuste a funciones no lineales. Puede sufrir con valores atípicos.

Interpretación del modelo:

OLS Regression Results

Dep. Variable:	Num_viajes	R-squared:	0.759
Model:	OLS	Adj. R-squared:	0.733
Method:	Least Squares	F-statistic:	29.17
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	5.45e-11
Time:	20:12:52	Log-Likelihood:	-480.40
No. Observations:	42	AIC:	970.8
Df Residuals:	37	BIC:	979.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.116e+04	6602.803	1.690	0.099	-2221.256	2.45e+04
Num_escuelas	2.1012	0.608	3.454	0.001	0.869	3.334
Num_hospitales	50.4021	14.839	3.397	0.002	20.336	80.468
Num_iglesias	431.7600	162.204	2.662	0.011	103.104	760.416
Num_zonas_tipicas	-448.5378	156.228	-2.871	0.007	-765.085	-131.990

Omnibus:	5.674	Durbin-Watson:	2.126
Prob(Omnibus):	0.059	Jarque-Bera (JB):	4.374
Skew:	0.650	Prob(JB):	0.112
Kurtosis:	3.899	Cond. No.	2.79e+04

Al analizar los resultados de la regresión lineal podemos ver que el impacto que hace el intercepto (la constante) no es estadísticamente significativo al tener un P value de 0.09, superando el margen de tolerancia de 0.05.

Asímismo vemos que el número de escuelas es la variable que tiene una mayor significancia estadística, ya que su t value es mayor al de las demás variables. Esta variable se relaciona de manera positiva con el número de viajes recibido. Es decir que a mayor escuelas, mayor cantidad de viajes recibidos en una comuna. Por cada escuela nueva que haya, se generarán 2 nuevos viajes.

Con respecto a los hospitales, vemos que es estadísticamente significativo (p value menor a 0.05). Esta variable se relaciona positivamente con la cantidad de viajes recibidos en una comuna, es decir que a mayor cantidad de hospitales, mayor cantidad de viajes recibidos. Por cada nuevo hospital que haya, se generarán 50 nuevos viajes.

Hablando de las iglesias, esta variable es estadísticamente significativa. Se relaciona positivamente con los viajes recibidos por comuna, por lo que a mayor

cantidad de iglesias, más viajes se reciben en una comuna. Por cada nueva iglesia que haya, se generarán 431 nuevos viajes.

En lo que respecta a las zonas típicas, la variable es estadísticamente significativa. Tiene una relación negativa con la cantidad de viajes, así que mientras más zonas típicas se tengan, menos viajes se recibe en la comuna. Por cada nueva zona típica, se reciben 448 viajes menos.

Evaluación del modelo:
El modelo obtuvo un MAPE del 67.03% al evaluar los datos de pruebas, este resultado está altamente alejado de nuestro objetivo de minería de datos, pero dando los resultados esperados previos al entrenamiento del modelo.

Modelo 2. Regresión Lineal

Debido a que se realizó una añadidura en la cantidad de variables, se utilizará nuevamente el modelo inicial de regresión lineal con el fin de saber si las variables añadidas son útiles para la mejora del modelo.

Framework: Sci-kit Learn
Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Código Fuente

Parámetros:
Al ser un modelo de regresión lineal y solo se busca ver si las variables agregadas mejoran el modelo, éste utiliza los parámetros por defecto de la librería.

Descripción del modelo:
Se espera una mejora de al menos un 10% con respecto al modelo anterior.

Al igual que en el modelo anterior, no se pueden modelar relaciones complejas. No se pueden capturar relaciones no lineales sin transformar la entrada, por lo que se tiene que trabajar duro para que se ajuste a funciones no lineales. Puede sufrir con valores atípicos.

Interpretación del modelo

Dep. Variable:	Num_viajes	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.760
Method:	Least Squares	F-statistic:	120.8
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	3.45e-291
Time:	19:04:12	Log-Likelihood:	-8547.5
No. Observations:	1020	AIC:	1.715e+04
Df Residuals:	992	BIC:	1.729e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1837.0358	169.705	-10.825	0.000	-2170.058	-1504.013
Num_escuelas	3.9580	0.232	17.069	0.000	3.503	4.413
Num_hospitales	102.5404	5.543	18.499	0.000	91.663	113.418
Num_iglesias	832.9501	61.073	13.639	0.000	713.103	952.797
Num_zonas_tipicas	-877.6988	58.570	-14.986	0.000	-992.633	-762.764
Hora_llegada_1	1549.4546	226.744	6.834	0.000	1104.502	1994.407
Hora_llegada_2	1813.1202	229.522	7.900	0.000	1362.717	2263.524
Hora_llegada_3	1146.8853	225.522	5.085	0.000	704.331	1589.440
Hora_llegada_4	818.8051	230.987	3.545	0.000	365.527	1272.083
Hora_llegada_5	526.5857	228.115	2.308	0.021	78.943	974.228
Hora_llegada_6	573.7907	234.032	2.452	0.014	114.537	1033.045
Hora_llegada_7	416.2766	229.509	1.814	0.070	-34.102	866.656
Hora_llegada_8	629.2827	225.469	2.791	0.005	186.831	1071.734
Hora_llegada_9	802.0096	226.748	3.537	0.000	357.049	1246.970
Hora_llegada_10	1272.2533	234.092	5.435	0.000	812.881	1731.626
Hora_llegada_11	1489.8534	226.737	6.571	0.000	1044.915	1934.792
Hora_llegada_12	2010.5730	223.057	9.014	0.000	1572.855	2448.291
Hora_llegada_13	2391.1551	224.257	10.663	0.000	1951.083	2831.227
Hora_llegada_14	2498.9646	226.737	11.021	0.000	2054.026	2943.903
Hora_llegada_15	2364.3889	234.053	10.102	0.000	1905.092	2823.686
Hora_llegada_16	2357.5997	226.737	10.398	0.000	1912.660	2802.539
Hora_llegada_17	2315.7669	232.417	9.964	0.000	1859.682	2771.852
Hora_llegada_18	2546.6235	226.749	11.231	0.000	2101.662	2991.585
Hora_llegada_19	2824.0621	232.476	12.148	0.000	2367.860	3280.264
Hora_llegada_20	3249.4284	235.626	13.791	0.000	2787.045	3711.812
Hora_llegada_21	3810.2181	232.399	16.395	0.000	3354.169	4266.267
Hora_llegada_22	2993.9341	234.112	12.788	0.000	2534.523	3453.346
Hora_llegada_23	2120.6275	224.272	9.456	0.000	1680.525	2560.730

Al analizar los resultados de la regresión lineal podemos ver que el impacto que hace el intercepto (la constante) ahora sí es estadísticamente significativo y tiene una alta magnitud.

Asímismo vemos que el número de escuela se relaciona de manera positiva con el número de viajes recibido. Es decir que a mayor escuelas, mayor cantidad de viajes recibidos en una comuna. Por cada nueva escuela, se reciben 3 viajes más.

Con respecto a los hospitales, vemos que es estadísticamente significativo (p value menor a 0.05), y es la variable más significativa de todas al tener el t value más alto. Esta variable se relaciona positivamente con la cantidad de viajes recibidos en una comuna, es decir que a mayor cantidad de hospitales, mayor cantidad de viajes recibidos. Por cada hospital nuevo, se reciben 102 viajes más.

Hablando de las iglesias, esta variable es estadísticamente significativa. Se relaciona positivamente con los viajes recibidos por comuna, por lo que a mayor cantidad de iglesias, más viajes se reciben en una comuna. Por cada nueva iglesia, se reciben 832 viajes más.

En lo que respecta a las zonas típicas, esta variable es estadísticamente significativa. Tiene una relación negativa con la cantidad de viajes, así que mientras más zonas típicas se tengan, menos viajes se recibe en la comuna. Por cada zona típica, se reciben 877 viajes menos.

Con respecto a las variables dummies, vemos que todas ellas son estadísticamente significativas para nuestro modelo, además de que tienen magnitudes mucho más altas que el resto de variables, por lo que sabemos que el cambio en la hora tiene un impacto más grande en la cantidad de viajes recibidos que las demás variables.

Evaluación del modelo:
El modelo obtuvo un MAPE del 54.92% al evaluar los datos de pruebas, por lo cual no cumple con nuestro objetivo de minería de datos. A comparación del modelo anterior se obtuvo una mejora del 12%.

Comparación modelos

Modelo	Hiperparámetros / Configuración	Score Train (%MAPE)	Score Test (%MAPE)
Regresión Lineal (1)	4 Variables	31.47%	67.03%
Regresión Lineal (2)	27 Variables (23 dummies)	44.04%	54.92%

Modelo 3. Regresión lineal

Debido al cambio de objetivo de negocio y de minería de datos se tiene que realizar un nuevo modelo benchmark, escogiendo nuevamente el modelo de regresión lineal por su simpleza para la interpretación.

Framework: Sci-kit Learn

Código fuente

Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Parámetros:
Como modelo benchmark se utilizan los parámetros por defecto de la librería.

Descripción del modelo:
Para este modelo se tiene esperado una variabilidad mayor a un 50%.

Recordando nuevamente las limitaciones como lo es no poder modelar relaciones complejas. No se pueden capturar relaciones no lineales sin transformar la entrada, por lo que se tiene que trabajar duro para que se ajuste a funciones no lineales. Puede sufrir con valores atípicos.

Prueba de multicolinearidad:
Al encontrarnos en un contexto de regresión lineal nos interesa averiguar si existen problemas de multicolinearidad en nuestras variables. Para averiguar esto, se ejecutarpa una prueba en la que se calculará el VIF de las variables. Aquellas con un VIF superior a 5 serán descartadas para este modelo en particular.

Resultado de la prueba:

	Variable	VIF
1	Num_escuelas	2.179267
2	Num_hospitales	4.861675
3	Num_iglesias	20.398728
4	Num_zonas_tipicas	28.342396
5	Hora_llegada_1	1.892614
6	Hora_llegada_2	1.854917
7	Hora_llegada_3	1.912862
8	Hora_llegada_4	1.835817
9	Hora_llegada_5	1.873951
10	Hora_llegada_6	1.796272
11	Hora_llegada_7	1.854715
12	Hora_llegada_8	1.911969
13	Hora_llegada_9	1.892683
14	Hora_llegada_10	1.797202
15	Hora_llegada_11	1.892496
16	Hora_llegada_12	1.950441
17	Hora_llegada_13	1.931511
18	Hora_llegada_14	1.892500
19	Hora_llegada_15	1.796607
20	Hora_llegada_16	1.892508
21	Hora_llegada_17	1.815142
22	Hora_llegada_18	1.892697
23	Hora_llegada_19	1.816073
24	Hora_llegada_20	1.775957
25	Hora_llegada_21	1.814857
26	Hora_llegada_22	1.797506
27	Hora_llegada_23	1.931779

Vemos como resultado de la prueba que las variables dummies no presentan el problema de multicolinearidad. Sin embargo, las variables de número de iglesias y zonas típicas tienen un VIF bastante superior a 5, por lo que las eliminaremos del modelo de regresión lineal.

Interpretación del modelo:

OLS Regression Results

Dep. Variable:	Num_viajes	R-squared:	0.720
Model:	OLS	Adj. R-squared:	0.713
Method:	Least Squares	F-statistic:	102.4
Date:	Tue, 29 Nov 2022	Prob (F-statistic):	1.72e-254
Time:	20:19:21	Log-Likelihood:	-7463.8
No. Observations:	1020	AIC:	1.498e+04
Df Residuals:	994	BIC:	1.511e+04
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-502.5967	57.773	-8.699	0.000	-615.968	-389.225
Num_escuelas	1.1840	0.064	18.479	0.000	1.058	1.310
Num_hospitales	15.3114	1.026	14.927	0.000	13.298	17.324
Hora_llegada_1	449.5168	78.280	5.742	0.000	295.903	603.131
Hora_llegada_2	581.1714	79.236	7.335	0.000	425.683	736.660
Hora_llegada_3	314.1786	77.843	4.036	0.000	161.424	466.933
Hora_llegada_4	238.1068	79.740	2.986	0.003	81.628	394.586
Hora_llegada_5	165.6759	78.743	2.104	0.036	11.154	320.198
Hora_llegada_6	224.0181	80.797	2.773	0.006	65.466	382.570
Hora_llegada_7	232.6279	79.217	2.937	0.003	77.177	388.079
Hora_llegada_8	294.7687	77.839	3.787	0.000	142.020	447.517
Hora_llegada_9	330.8361	78.280	4.226	0.000	177.223	484.449
Hora_llegada_10	424.8162	80.784	5.259	0.000	266.290	583.342
Hora_llegada_11	527.5747	78.279	6.740	0.000	373.964	681.185
Hora_llegada_12	721.5535	77.007	9.370	0.000	570.439	872.668
Hora_llegada_13	851.1410	77.416	10.994	0.000	699.223	1003.059
Hora_llegada_14	864.2274	78.279	11.040	0.000	710.617	1017.838
Hora_llegada_15	772.1980	80.784	9.559	0.000	613.671	930.725
Hora_llegada_16	790.9504	78.278	10.104	0.000	637.341	944.560
Hora_llegada_17	858.6018	80.239	10.701	0.000	701.145	1016.058
Hora_llegada_18	967.7696	78.280	12.363	0.000	814.157	1121.382
Hora_llegada_19	1158.6340	80.247	14.438	0.000	1001.161	1316.107
Hora_llegada_20	1433.9761	81.347	17.628	0.000	1274.344	1593.608
Hora_llegada_21	1800.0873	80.234	22.436	0.000	1642.641	1957.534
Hora_llegada_22	1281.6126	80.817	15.858	0.000	1123.021	1440.204
Hora_llegada_23	833.9316	77.424	10.771	0.000	681.998	985.865

Al analizar los resultados de la regresión lineal podemos ver que el impacto que hace el intercepto (la constante) ahora sí es estadísticamente significativo y tiene una alta magnitud.

Asímismo vemos que el número de escuela se relaciona de manera positiva con el número de viajes recibido. Es decir que a mayor escuelas, mayor cantidad de viajes recibidos en una comuna. Por cada nueva escuela, se recibe un viaje más.

Con respecto a los hospitales, vemos que es estadísticamente significativo (p value menor a 0.05), y es la variable más significativa de todas al tener el t value más alto. Esta variable se relaciona positivamente con la cantidad de viajes recibidos en una comuna, es decir que a mayor cantidad de hospitales, mayor cantidad de viajes recibidos. Por cada hospital nuevo, se reciben 15 viajes más.

Con respecto a las variables dummies, vemos que todas ellas son estadísticamente significativas para nuestro modelo, además de que tienen magnitudes mucho más altas que el resto de variables, por lo que sabemos que el cambio en la hora tiene un impacto más grande en la cantidad de viajes recibidos que las demás variables. También vemos que las horas de más impacto son aquellas en el rango de las 19hrs a las 22 hrs.

Evaluación del modelo:
El modelo obtuvo un MAPE del 87.8% al evaluar los datos de pruebas, este resultado está altamente alejado de nuestro objetivo de minería de datos. Este modelo brinda los resultados esperados al momento de su elección como modelo benchmark.

Modelo 4. Random forest

Los modelos de decisión basados en bosques (conjuntos de árboles) se caracterizan por ser más precisos que los modelos sencillos generados a base de regresiones lineales. Además, estos modelos nos permiten conocer la significancia de cada variable que integra al modelo, por lo que nos resultará de gran utilidad para cumplir con el objetivo de negocio de conocer cuales son los principales atractores de viajes.

Framework: Sci-kit Learn
Código Fuente

Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Parámetros:
Debido a que este modelo se utilizará únicamente como un paso intermedio en la mejora del rendimiento, se utilizarán los valores por default del regresor de random forest de sklearn, es decir:

- n_estimators: 100
- max_depth : None
- min_samples: None

Descripción del modelo:

Para este modelo se tiene esperado una variabilidad alrededor de un 30%.

Las desventajas de este modelo es que se pierde la facilidad de interpretación del modelo, además de que no puede predecir más allá del rango de valores del conjunto de entrenamiento.

Evaluación del modelo:
El modelo obtuvo una variación del 15.05% al evaluar los datos de pruebas, el resultado es casi aceptable para el objetivo de mineria,

Los resultados obtenidos son muy superiores a los esperados, el modelo superó al anterior en un 74% siendo una mejora bastante significativa.

Modelo 5. XGBoost

El modelo XGBoost nos ofrece un rendimiento superior a los modelos de árboles tradicionales gracias a que implementa una optimización avanzada utilizando Gradient Boosting.

Framework: Sci-kit Learn
Código Fuente

Este modelo recibe como variables independientes lo siguiente:
- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:
- Número de viajes hacia la comuna

Parámetros:
Para afinar los hiperparámetros de este modelo, recurrimos a utilizar una búsqueda en cuadrícula CV. Este enfoque nos permite probar diferentes valores de los hiperparámetros para encontrar los que ofrecen el mejor rendimiento.

Dándonos como resultado los siguientes hiperparametros:

- colsample_bytree: 0.1
- learning_rate: 0.1
- max_depth: 3
- n_estimators: 20000

Descripción del modelo:

Para este modelo se tiene esperado una variabilidad menor al 20%

Las limitaciones de este modelo es que sus resultados pueden llegar a ser complejos de interpretar, puede llegar a ajustar ciertos grupos de datos en presencia de ruido y se tiene poco control sobre lo que hace el modelo.

Evaluación del modelo:
El modelo obtuvo un MAPE del 14.84 % al evaluar los datos de pruebas, el resultado cumple con el objetivo de la minería de datos.

Modelo 6. Red Neuronal MLP

El modelo Multi-layer Perceptron regressor nos permite hacer una regresión a través de una red neuronal usando optimizadores basados en gradient descent.

Framework: Sci-kit Learn
Código Fuente

Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Parámetros:
Para la selección de parámetros se realizaron múltples pruebas con los distintos tipos de activación para encontrar el modelo que ofrezca mejores resultados. Después de las pruebas, hallamos buenos resultados con los siguientes hiperparámetros.

- random_state=1,
- max_iter=1000000,
- learning_rate="adaptive",
- activation = "logistic"

Descripción del modelo:

Para este modelo se espera obtener un resultado igual o mejor al de XGBoost, es decir, un 20% o menos de error.

Evaluación del modelo:
Este modelo presentó un MAPE de 67% al evaluar los datos de pruebas. Creemos que esto se debe a que el tamaño de nuestro dataset de entrenamiento no es lo suficientemente grande como para alimentar un modelo profundo.
Debido al resultado, se decidió no utilizar técnicas de Deep Learning para la creación del modelo. En su lugar, nos apegaremos al modelo XGBoost, ya que nos brindó los mejores resultados.

Comparación de modelos

Modelo	Hiperparámetros/ Configuración	Score Train (%MAPE)	Score Test (%MAPE)
Regresión Lineal	• 27 Variables (23 dummies)	63.55%	87.80%
Random Forest	• 27 Variables (23 dummies) • n_estimators: 100 • max_depth : None • min_samples: None	14.98%	15.05%
XGBoost	• 27 Variables (23 dummies) • colsample_bytree: 0.1 • learning_rate: 0.1 • max_depth: 3 • n_estimators: 20000	10.72%	14.89%
Red neuronal MLP	• 27 Variables (23 dummies) • random_state=1, • max_iter=1000000, • learning_rate="adaptive", • activation = "logistic"	58.31%	67.23%

Modelo 7. XGBoost

Este modelo se basa en el modelo 5, con la diferencia de que se construyó un nuevo dataset de entrenamiento que incluye un mayor número de variables. Esto se hizo basado en la sugerencia de nuestro SF.

Framework: Sci-kit Learn
Código Fuente

Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- Superficie en Km^2
- Población
- Conexiones fijas de internet
- M^2 de areas verdes
- Número de micro empresas
- Número de empresas pequeñas
- Número de empresas medianas
- Número de empresas grandes
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Descripción del modelo:
Para este modelo se tiene esperado una variabilidad menor al 20%

Las limitaciones de este modelo es que sus resultados pueden llegar a ser complejos de interpretar, puede llegar a ajustar ciertos grupos de datos en presencia de ruido y se tiene poco control sobre lo que hace el modelo.

Parámetros:
Como parámetros iniciales para el modelo xgboost utilizaremos los valores por default del framework sklearn para xgboost. El único valor que modificaremos es el de objective. Indicamos al modelo que debe utilizar como objetivo una regresión de Poisson, ya que estas regresiones están orientadas a predecir el conteo de valores y únicamente generan resultados positivos, a diferencia de la regresión lineal clásica que puede generar valores negativos.

Esta primera versión nos da un % de error (MAPE) de **27.4%** en train y **31.24%** en test. Debido a que estos valores están por detrás de lo visto en nuestros modelos anteriores, procedemos a realizar un refinamiento.

Primer refinamiento:
Para afinar los hiperparámetros de este modelo, recurrimos a utilizar una búsqueda en cuadrícula CV. Este algoritmo realiza una fuerza por medio de fuerza bruta, probando con un conjunto de hiper parámetros para evaluar cuáles otorgan mejores resultados.

Al aplicar este algoritmo, obtuvimos como resultado los siguientes hiperparametros:

- colsample_bytree: 0.1
- learning_rate: 0.1
- max_depth: 3
- n_estimators: 100000

Con la optimización de gridSearchCV logramos obtener un % de error (MAPE) de **1%** en train y **14.1%** en test. Si bien el error de test es bastante bueno, vemos que hay overfitting, por lo que no es un modelo muy estable. Debido a esto, decidimos hacer de forma manual un nuevo refinamiento del modelo.

Segundo refinamiento:
Para este segundo refinamiento ajustamos de manera manual los hiperparámetros. Se tomaron los valores del primer refinamiento pero se redujo la cantidad de estimadores. Esto nos permitirá reducir la posibilidad de que el modelo memorice los valores de entrenamiento, generando un modelo más estable.

- colsample_bytree: 0.1
- learning_rate: 0.1
- max_depth: 3
- n_estimators: 10000

Con el refinamiento manual al modelo, obtuvimos un % de error (MAPE) de **10.4%** en train y **16.9%** en test. De esta manera, logramos reducir en una gran medida el overfitting sin tener un impacto muy grande en la precisión del modelo.

Cross fold validation
Para este modelo se realizó una validación cruzada de K-folds con un valor de 10 folds. Se eligió esta cantidad de folds ya que, como se comenta en el artículo “Why use cross fold validation” de KDNuggets, al usar 10 folds se suele tener un bias bajo y una varianza moderada en los resultados, representando muchas veces un punto óptimo.
El resultado obtenido como precisión es de: 85.78%

```
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score

kfold = KFold(n_splits=10)
results = cross_val_score(xgb_model, X, y, cv=kfold)
print("Accuracy: %.2f%% " % (results.mean()*100))
```

```
Accuracy: 85.78%
```

Evaluación del modelo:

Modelo	Hiperparámetros / Configuración	Score Train (%MAPE)	Score Test (%MAPE)
Modelo Inicial	34 Variables (23 dummies) ● default sklearn values	27.4%	31.2%
Primer Refinamiento	34 Variables (23 dummies) ● colsample_bytree: 0.1 ● learning_rate: 0.1 ● max_depth: 3 ● n_estimators: 100000	1%	14.1%
Segundo Refinamiento	34 Variables (23 dummies) ● colsample_bytree: 0.1 ● learning_rate: 0.1 ● max_depth: 3 ● n_estimators: 10000	10.4%	16.9%

El modelo en su segundo refinamiento obtuvo una variación del 16.98% al evaluar los datos de pruebas, el resultado cumple con el objetivo de la minería de datos.
Es importante mencionar que, si bien este modelo obtuvo una precisión inferior al primer refinamiento, nos aporta un mayor valor tener un modelo más estable gracias a la reducción del overfitting.

Bibliografia

[1]Kim, S.; Kim, H.(2016) A new metric of absolute percentage error for intermittent demand forecasts. Recuperado de https://reader.elsevier.com/reader/sd/pii/S0169207016000121?token=79EECEB30F7CC2154434EABF95C521C124BE6FA065E66ADC0D2DA62069333CAB8F055338228D91B26392CF3C0A3843AF&originRegion=us-east-1&originCreation=20221128202145

[2]Swanson, D.A.; Tayman, J.; Bryan, T.M.(2011) MAPE-R: A RESCALED MEASURE OF ACCURACY FOR CROSS-SECTIONAL, SUBNATIONAL FORECASTS. Recuperado de https://paa2011.populationassociation.org/papers/110062

[3]De Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F.(2017, julio 10) Mean Absolute Percentage Error for regression models. Recuperado de https://arxiv.org/abs/1605.02541