Caso: Clasificación de Oportunidades de Venta para Empresa de Refrigeración

Manuel Contreras
Analista en Tenaris
Ing. Industrial UTN

De Leo, Andrés
Pasante Trade Marketing en Bayer
Ing. Industrial UTN

Tondini, Matías
Analista Jr. En Danone
Ing. Industrial UTN

***Abstract*—El objetivo del trabajo es lograr determinar la probabilidad de como cerrará un caso, ya sea como "Close Won" o "Close Lost"**

Close won— Oportunidades comerciales cerradas como ventas efectivas.

Close lost— Oportunidades comerciales cerradas sin lograr vender.

I. INTRODUCCIÓN

El trabajo se basa en un complejo análisis multivariado en el cual el fin es lograr predecir si un caso concretará en una venta (Close Won) o se cerrará como caso perdido (Close Lost)

El caso consiste en un negocio de venta de refrigerante de una empresa B2B ("Business to Business"), es esencial para ellos optimizar los esfuerzos de los representantes comerciales, ayudándolos a priorizar las oportunidades en el pipeline. Una oportunidad consiste en un proyecto de venta o instalación de equipos para un cliente.

Para ellos contamos con diferentes variables, entre ellas las más importantes: el vendedor a cargo de la venta, la cantidad de toneladas de refrigerante (TRF), fecha prevista de entrega de los equipos, información geográfica de los clientes, el precio, entre otras.

Con este dato la empresa lograría que sus ejecutivos comerciales lograran enfocar sus esfuerzos en las oportunidades de venta más probables de hacerse efectivas.

II. DATA

A. Fuente

El set de datos proviene de datos históricos de la empresa Alix Partners.

El mismo contiene 30 variables categóricas, 7 tipo fecha y 15 numéricas.

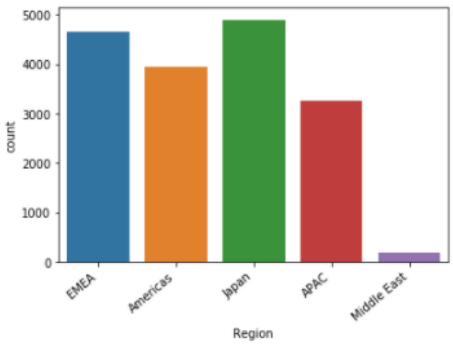
B. EDA

Una de las problemáticas más importantes de este caso fue lograr una correcta limpieza de los datos.

En ellos encontramos los siguientes problemas:

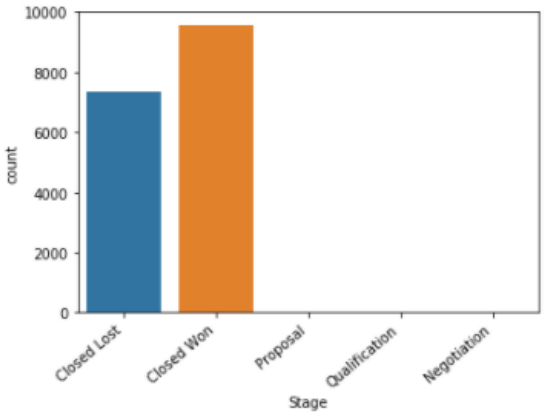
- Diferencia en las unidades monetarias de cada región, por lo que tuvimos que unificarlas.
- Muchas variables categóricas, las cuales convertimos en dummies, entre ellas, las más importantes fueron el vendedor a cargo de la oportunidad y el tipo de caso (Won o Lost).
- Campos vacíos en gran cantidad de columnas.
- Columnas con datos que no agregaban valor al análisis, las cuales fueron eliminadas.

En primer lugar, para tener una noción de cómo se distribuyen las ventas por región realizamos un countplot **(figura 1)**



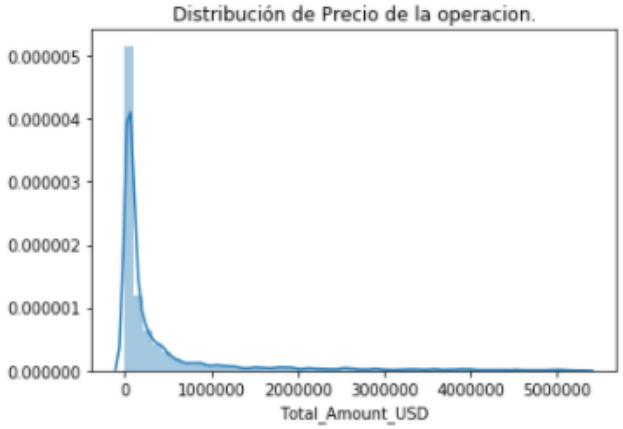
(Figura 1)

Verificamos la cantidad de casos que hay de cada estado final (figura 2)



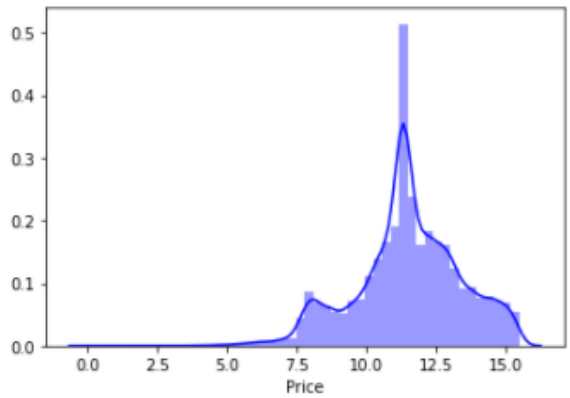
(Figura 2)

Luego, con un distplot buscamos poder visualizar el rango de los precios totales de operación más frecuentes **(Figura 3)**



(Figura 3)

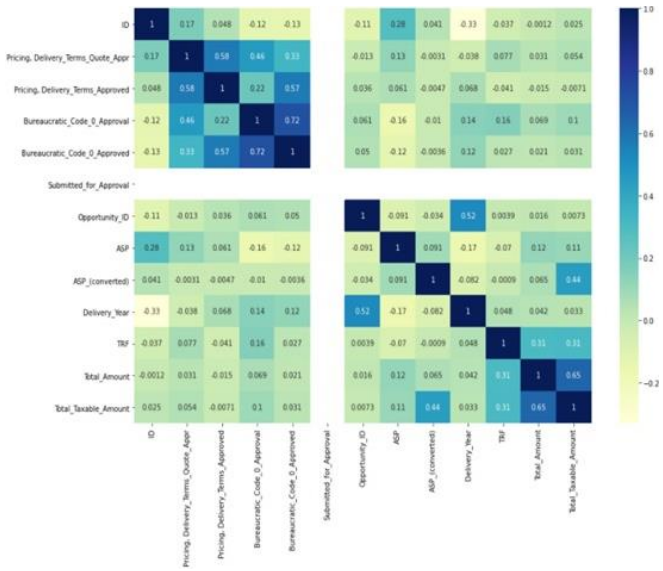
Debido a que se puede apreciar muy poco en este primer intento de visualizar la distribución de precios totales, realizamos nuevamente un distplot pero esta vez afectándolo por escala logarítmica como puede verse en la **figura 4:**



(figura 4)

Como puede apreciarse en esta última visualización la media de los precios se encuentra en 11,6 USD y la gran mayoría en un rango entre 7,5 y 15,5 USD.

Para finalizar con el EDA buscamos aquellas variables que podrían tener cierta correlación lineal entre ellas. Para ello, realizamos un heatmap en una matriz de correlación de Pearson:



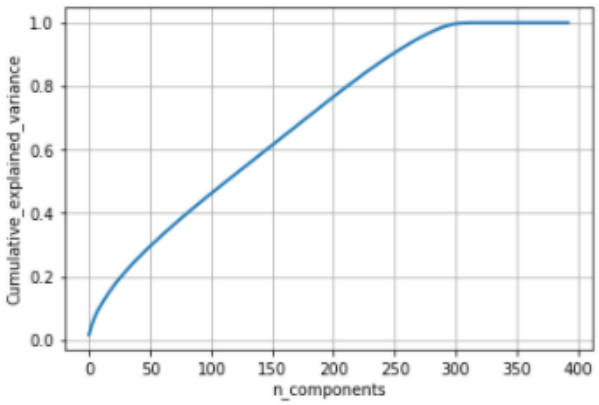
(Figura 7)

C. Análisis de Componentes Primarios (PCA)

Luego de todo lo que correspondió a la limpieza de datos, obtuvimos una base de datos reducida, pero que debido a la gran cantidad de variables categóricas que para nuestro entender eran de importancia para el proyecto, se recurrió a la técnica de la generación de Dummies.

Esto consiste en generar nuevas columnas que reemplacen a las existentes, pero que, por cada registro único en cada una de las columnas originales, se creará una nueva feature que tendrá en sus valores 0 ó 1 según corresponda. Habrá un 0 si la sample no tiene tal atributo y habrá un 1 si lo tiene.

En consecuencia, el dataset aumentó su cantidad de columnas a 394. Por tal motivo, se tomó la decisión de realizar un PCA. Este método tiene como objetivo el de reducir la dimensionalidad del set de datos, quedándose con las variables que más información aporten al modelo. [1]



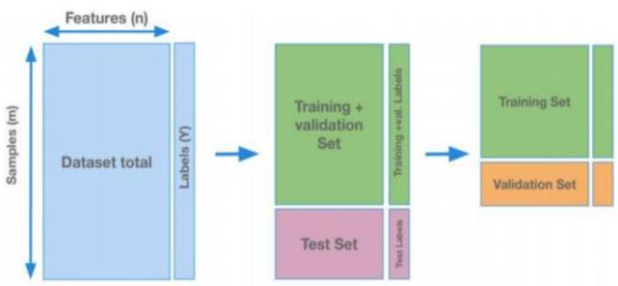
(Figura 8)

Como puede verse en la figura 6, el total de la variabilidad puede explicarse con tan sólo 294 features. Pero como contrapartida a esto, ninguna de ellas es realmente significativa, ya que la que mejor explica el modelo tiene menos del 2% de la variabilidad.

III. MODELOS

A. Explicación

En primer lugar, se separó la variable a predecir del set de datos. Luego, se dividió el dataset en Train y Test. Este método se realiza con el fin de entrenar a nuestros modelos con una parte de los datos y luego realizar una prueba con nuevos datos desconocidos para ellos y así, poder identificar si son capaces de clasificarlos correctamente.



Además, para identificar los mejores hiper-parámetros para los modelos propuestos se utilizó el método GridSearch que permitió probar distintos hiper-parámetros

Por otro lado, también se implementó una técnica llamada StandardScaler, que tiene como objetivo final el estandarizado de las medias y desvíos de todos los atributos. Luego de aplicado este escalado de datos, la media de todas las features quedaron con valor 0 y sus desvíos en 1.

Una vez hecha la división en train y test, la estandarización de las variables y encontrados los mejores hiper-parámetros para cada modelo, se prosiguió a poner en funcionamiento cada uno de ellos. Para este caso en particular escogimos los siguientes:

- Random Forest: Este método genera distintos árboles de decisión según posibles caminos que se forman al ir cumpliendo determinadas series de condiciones lógicas. [2]
- KNN: el modelo consiste en determinar los K vecinos más cercanos por distancia euclídea (distancia par a par). [3]
- Logistic Regresion: Es un clasificador lineal, que consiste en una Regresión Lineal, precedida de una función Sigmoide. El fin de esta Regresión es asignarle una probabilidad a cada muestra de pertenecer a una u otra clase.

Para poder penalizar las muestras mal clasificadas dimensionamos al hiper-parámetro C, que les asigna un valor de penalización a cada muestra errónea. A menor valor de C menos complejo será el modelo, mientras que si C aumenta el modelo será de una geometría compleja en la cual cada vez se ajuste mejor a clasificar todos los datos. Esto puede traer como contrapartida un sobre ajuste en el cual se clasifiquen demasiado bien los datos de entrenamiento, pero que, al testearlo con nuevos datos, el modelo será proclive a tener un nivel de Accuracy bajo. [4]

- Support Vector Classifier (SVC): Es un clasificador lineal, que busca el hiperplano separador que maximiza el margen entre las clases. Al igual que en Logistic Regression trabaja con un penalizador C para ajustar el modelo.

Para evaluar los modelos podemos resumirlos en la siguiente tabla, con 2 métricas la cuales nos dan una referencia de que tan acertados son para clasificar nuevos casos.

Finalmente, logramos los siguientes resultados:

	Modelo	Accuracy	AUC
0	KNN_reducido	0.7728	0.7695
1	Random Forest	0.8597	0.856
2	KNN	0.777	0.7695
3	Logistic Regression	0.761	0.7514
4	SVC	0.766	0.7565

B. Matrices de confusión:

Con el objetivo de obtener un mejor entendimiento de los resultados, se decidió realizar una matriz de confusión para el Random Forest, el KNN y el SVC. Esta matríz permite divisar los valores True Positive, True Negative, False Positive y False Negative que arrojaron cada uno de los modelos.

Random Forest:

[1611 215]
[239 1166]

Valores Positivos calculados correctamente (TP): 1611
Valores Positivos calculados como negativos (FN): 239
Valores Negativos calculados correctamente (TN):1166
Valores Negativos calculados como positivos (FP): 215

KNN:

[1521 305]
[413 992]

Valores Positivos calculados correctamente (TP): 1521
Valores Positivos calculados como negativos (FN): 413
Valores Negativos calculados correctamente (TN): 992
Valores Negativos calculados como positivos (FP): 305

SVC:

[1514 312]
[444 961]

Valores Positivos calculados correctamente (TP): 1514
Valores Positivos calculados como negativos (FN): 444
Valores Negativos calculados correctamente (TN): 961
Valores Negativos calculados como positivos (FP): 312

C. Conclusion

Como puede apreciarse en las últimas tablas, el modelo con mayor precisión el Random Forest con una probabilidad de acierto del 86%.

Podemos concluir que lo más complejo del trabajo fue hacer una exploración y correcta limpieza de los datos. Una vez con los datos listos para trabajar intentamos distintos modelos con el fin de encontrar el mejor estimador y lograr conocer cuando un caso podría llegar a concretarse de forma ganadora o perdedora para el vendedor.

Nos resultó interesante realizar este trabajo ya que un caso de probabilidad comercial es bastante abarcativo y, además, es un caso bastante genérico que nos sirvió para consolidar nuestros conocimientos en un caso práctico real.

Reconocimientos

Agradecemos al nuestro docente Martín Palazzo, nuestro mentor: Agustín Velázquez y a todo el equipo docente por su gran predisposición. Así como también reconocer el trabajo que dedicaron al brindarnos su tiempo en guiarnos frente a dudas y en preparar las clases para que sean lo más didácticas posibles. Principalmente hay que destacar que lograron un gran trabajo como equipo docente a pesar de la dificultad de adaptarse en poco tiempo a esta nueva modalidad de enseñanza virtual debido a la cuarentena que atraviesa por estos tiempos.

References

[1] Roweis, S. (1997). EM algorithms for PCA and SPCA. Advances in neural information processing systems, 10, 626-632.

[2] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222. K. Elissa, “Title of paper if known,” unpublished.

[3] Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. Computers & security, 21(5), 439-448.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[5] Tax, D. M., & Duin, R. P. (2004). Support vector data description. Machine learning, 54(1), 45-66.