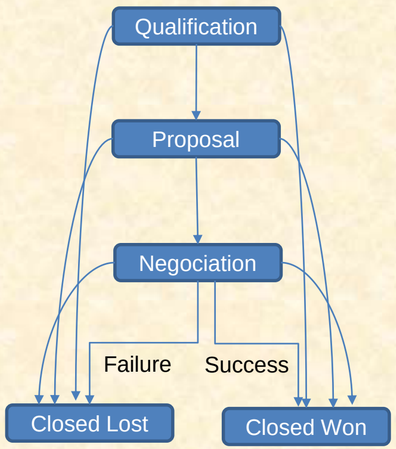
Clasificación de Oportunidades de Venta para Empresa de Refrigeración

Contreras, Manuel[1]; De Leo, Andrés[1] ; Tondini, Matías[1]

[1]Universidad Tecnologica Nacional, Facultad Regional Buenos Aires

Introducción

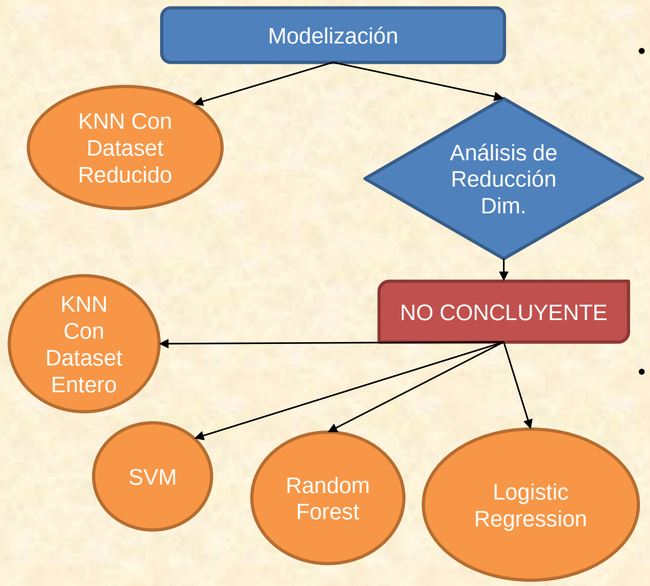


El presente trabajo práctico se encuentra formado por 3 partes: 1) Visualización de los datos, 2) limpieza 3) clasificación por medio de distintos modelos de Machine Learning.

El problema que se nos plantea se basa en poder identificar las **oportunidades de venta** para el negocio de la refrigeración de una empresa que se encarga de vender equipos para este fin. Para ello, propondremos distintos modelos de clasificación los cuales nos ayudaran a determinar aquellos casos que cumplen con las condiciones para que sean considerados “Closed WON” o “Closed LOST”. De esa manera podremos ayudar a la empresa a conocer de forma anticipada los casos en los que más posibilidades tienen de poder concretar la venta.

El tipo de aprendizaje de nuestros modelos será Supervisado, ya que tenemos información de las etiquetas de casos pasados.

Métodos



- En el Pipeline del proyecto, se puede divisar que propusimos distintos métodos de clasificación para el caso en cuestión. Planteamos dos métodos de KNN, uno con el dataset reducido por nosotros y otro con el dataset entero; un modelo de Support Vector Machines; otro Random Forest y por último un Logistic Regression.
- Además, intentamos reducir la dimensionalidad de nuestras Features para poder encontrar aquellas que mas variabilidad explicaban, pero no tuvimos éxito ya que ninguna de ellas superaba el 1%.

Resultados

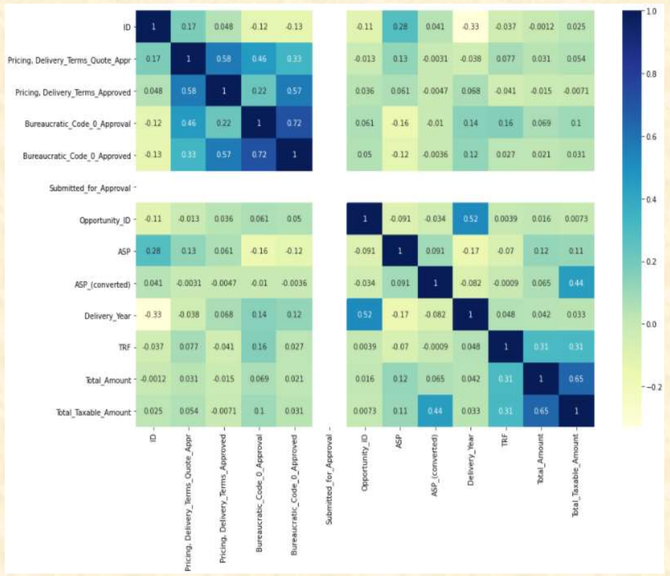
	Modelo	Accuracy	AUC
0	KNN_r	0.772826	0.769559
1	Random Forest	0.859796	0.856075
2	KNN	0.777778	0.769509
3	Logistic Regression	0.761065	0.751441
4	SVC	0.766017	0.756560

- Luego de aplicar todos los modelos mencionados, hicimos una tabla para poder determinar la precisión en la clasificación de cada uno de ellos, y vemos que el más destacado es el Random Forest con un 86% de Accuracy en la clasificación. El resto de los modelos oscila en valores cercanos al 77%.

Dataset

- La información fue brindada por la empresa Alix Partners.
- El Dataset cuenta con 16.947 samples y 52 features, con gran parte de ellas categóricas.
- Las Features que más se destacan a nuestro entender son:
- Territory
- Billing_Country
- Opportunity_Owner
- Planned_Delivery_Start_Date
- Planned_Delivery_End_Date
- Total_Amount
- Stage

Análisis Exploratorio de datos



- Al hacer una correlación lineal de Pearson entre las variables númericas , detectamos que ninguna de éstas explica de forma significativa a otra. Además, vemos que la Feature Submitted_For_Approval puede ser eliminada ya que no nos aporta nada de información.
- En el segundo gráfico, podemos ver la cantidad de oportunidades creadas por país. En primer lugar se encuentra Alemania con 1600 registros y le sigue Estados Unidos con 1500.
- El hecho de tener demasiadas variables categóricas nos obligó a la creación de nuevas variables Dummies para poder establecer relación entre ellas.

Conclusiones

- Según los resultados obtenidos, podemos decir que la tasa de aciertos de un 86% resulta convincente para poder trabajar con este modelo.
- Sin embargo, entendemos que este porcentaje podría verse mejorado en el futuro con una limpieza de datos más eficiente.