**Puntos:**

1. Introducción, objetivo y explicación.
2. Análisis exploratorio.
3. Importación de librerías y archivo.
4. Inspección de tabla y limpieza, tipos de datos.
5. Análisis de columnas y valores, imputación, eliminación.
6. Gráficos, agrupaciones descriptivas.
7. Reportes (nuevo, librería especial Sweetviz).
8. Feature Engineering (si es necesario).
9. Over Sampling o UnderSampling. Preguntar a Tomas cuál usar
10. Dummificación.
11. Clustering (ni idea)
12. Métrica a usar/mejorar para el proyecto. Justificación. F1, las otras de F1 (weighten macro, micro)?
13. Modelo KNN.
14. Modelo Naive Bayes (separar categóricas y numéricas, tienen tratamiento diferente).
15. Modelo Regresión Logística.
16. Modelo Árbol de decisión simple.
17. Modelo de ensamble.
18. Bagging
19. RandomForest
20. Boosting

* Gradient Boost
* ADA Boost
* XG Boost
* LightGBM

1. GridSearch en cada modelo.
2. Selección del mejor modelo.
3. Pipeline + GridSearch.
4. Feature importance. (Está bien hacerlo sobre un Boosting). <https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598>

Sharp

1. PCA o TSN?
2. APIS y modelo real en BBDD.

Observaciones generales:

1. Comentar el código: qué se hace en cada bloque
2. Posible conclusión de lo obtenido.

Elementos importantes:

* En problemas de Churn siempre estará **desequilibrado** en favor de los casos que no abandonan.
* Las **importancias de las características** se pueden extraer de nuestros modelos para comprobar qué características han contribuido más para que el modelo decida si un cliente va a abandonar o no. Podemos caer en la **maldición de la dimensionalidad**.
* Elegir la(s) métrica(s) que realmente refleja el aporte de valor del modelo al negocio es clave y tiene que estar claro antes de empezar el diseño del conjunto de datos y el modelo. ¡Las métricas son bastante buenas! Un acierto (accuracy) del 85%! ¡Fantástico! O… bueno, pensemos un poco antes. ¿Es realmente el accuracy la métrica que queremos observar? ¿Qué quiere detectar nuestro negocio con éste modelo?

Esa es la pregunta que debemos hacernos a la hora de evaluar un modelo. En este caso, la detección de aquellos casos que SÍ abandonan, queremos ver **Precisión** **(¿cuándo predice positivos, que porcentaje clasifica correctamente? Significa que quiero tener un bajo número de falsos positivos**) De esta forma estaremos haciendo uso correcto de los recursos ($) para retener al cliente. También nos importa el recall, así que podríamos usar **score F1**. Parece que algunas de las métricas que habitualmente se usan para evaluar modelos (accuracy) de éste tipo pueden ser engañosas.

Clase 33 – Evaluación de modelo: tiene bueno para mostrar los resultados del modelo (hacer tabla Excel parecida sino. Gráfico d Feature Importance de Árbol de ensamble y su tabla.

Conclusiones de otro trabajo:

CONCLUSIONES.

Existe un problema con el servicio de Internet por fibra óptica, que está haciendo que los clientes abandonen la compañía.

Derivado del problema con el servicio de internet, el resto de servicios dependientes de este, están viéndose afectados y hacen que el abandono aumente.

La compañía es capaz, en general, de mantener a los clientes a lo largo del tiempo. La antigüedad está ejerciendo un efecto negativo en el abandono.

Los clientes con un contrato de 1/2 años, abandonan la compañía mucho menos que los clientes con contratos mensuales. La compañía debe darle prioridad a este tipo de contratación.

Los clientes que eligen una forma de pago automática, abandonan la compañíaa en mucho menor medida que los clientes que no contratado una forma de pago automática. La compañía debe priorizar este tipo de formas de pago frente a las formas que requieren comprobación por parte del cliente.