Modelo predictivo de siniestros en Telecom S.A

Xgboost

People Analytics

CAPITAL HUMANO

Agenda de la presentación

- Propósito del desarrollo.
- Predecir vs explicar.
- Árbol de decisión
- Metodología utilizada.
- Evaluación y Resultados del modelo.
- ¿Por qué sabemos que funciona?
- Conclusiones

Propósito general:

Generar un modelo predictivo de accidentes que permita emprender acciones preventivas para evitarlos.

Propósitos específicos:

- Preservar la salud de los empleados de Telecom.
- Reducir los costos asociados a días caídos por accidentes de trabajo.
- Contribuir a mantener operativos los servicios afectados por empleados accidentados.

Problema:

Tomar un individuo de una población y clasificarlo como Hombre o Mujer

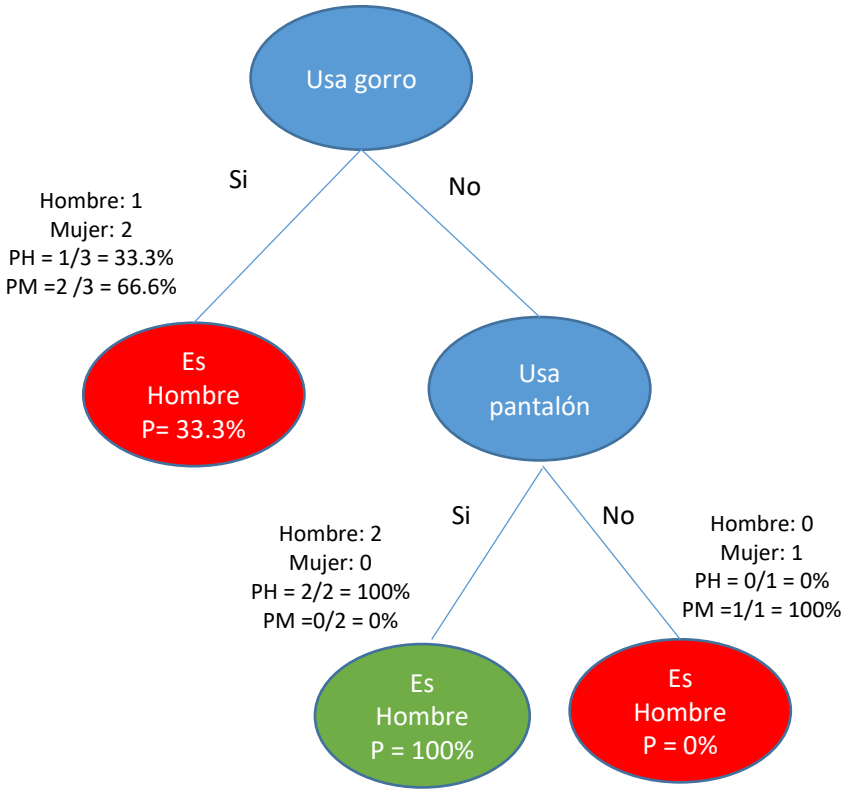
Potenciales variables predictoras:

- Largo del pelo.
- Bello facial
- Uso de gorros
- Uso de guantes
- Uso de maquillaje
- Uso de pantalones

Ninguna de estas variables permite explicar porqué un individuo es varón o mujer, sin embargo son extremadamente útiles para clasificarlos correctamente.

Una posible solución al problema anterior: Árbol de decisión

Para este problema imaginamos que solo contamos con 2 variables: “Usa gorro”, “Usa pantalón”



Datos conocidos

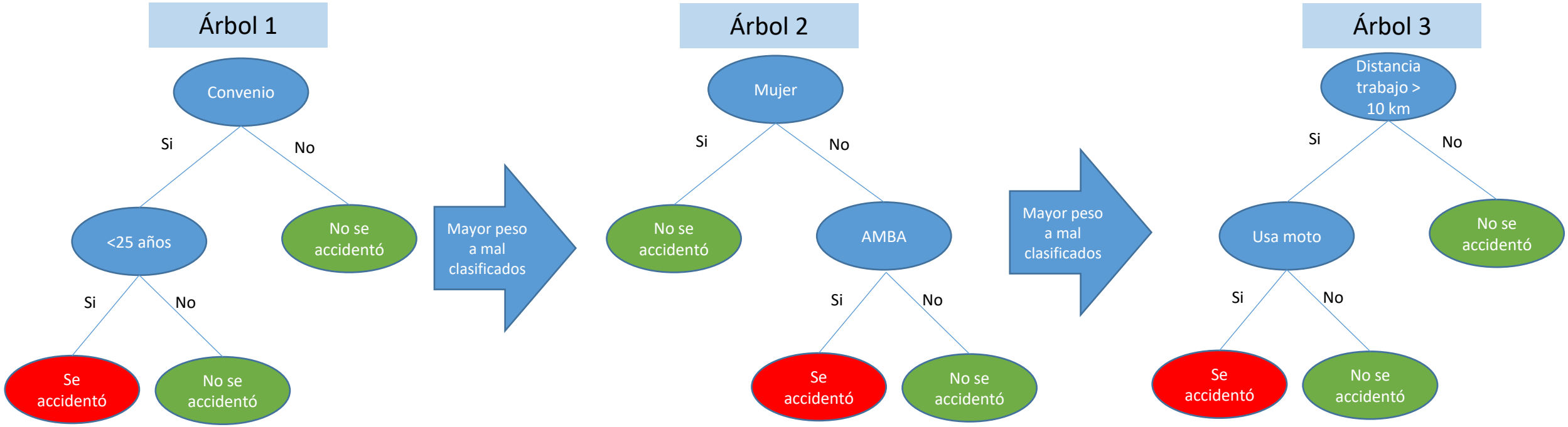
nro_individuo	Usa gorro	Usa pantalón	Es Hombre
1	si	si	no
2	si	si	no
3	no	no	no
4	no	si	si
5	no	si	si
6	si	si	si

Datos nuevos a clasificar

nro_individuo	Usa gorro	Usa pantalón	Es Hombre
7	no	si	?
8	no	no	?
9	no		?
10	si	si	?
11	no	si	?
12	si	si	?

Metodología utilizada:

Xgboost: Es un modelo de machine learning que suele ser utilizado en competencias de ciencia de datos y en la industria (ej. Cisco, Spotify, Novetta) por alcanzar los mas altos niveles de performance.



Evaluación y Resultados del modelo mes de marzo 2020:

- La tasa de acierto al utilizar el modelo está entorno al 2% de las predicciones.
- La tasa de aciertos haciendo predicciones al azar está entre el 0.4% y el 0.6% de las predicciones hechas. 0.4% para dotación completa y 0.6% para dotación en convenio
- El Modelo mejora entre 3 y 5 veces las predicciones hechas al azar.
- El testeo es "ácido" porque se muestran predicciones en un mes con menos casos de los que usualmente suceden.

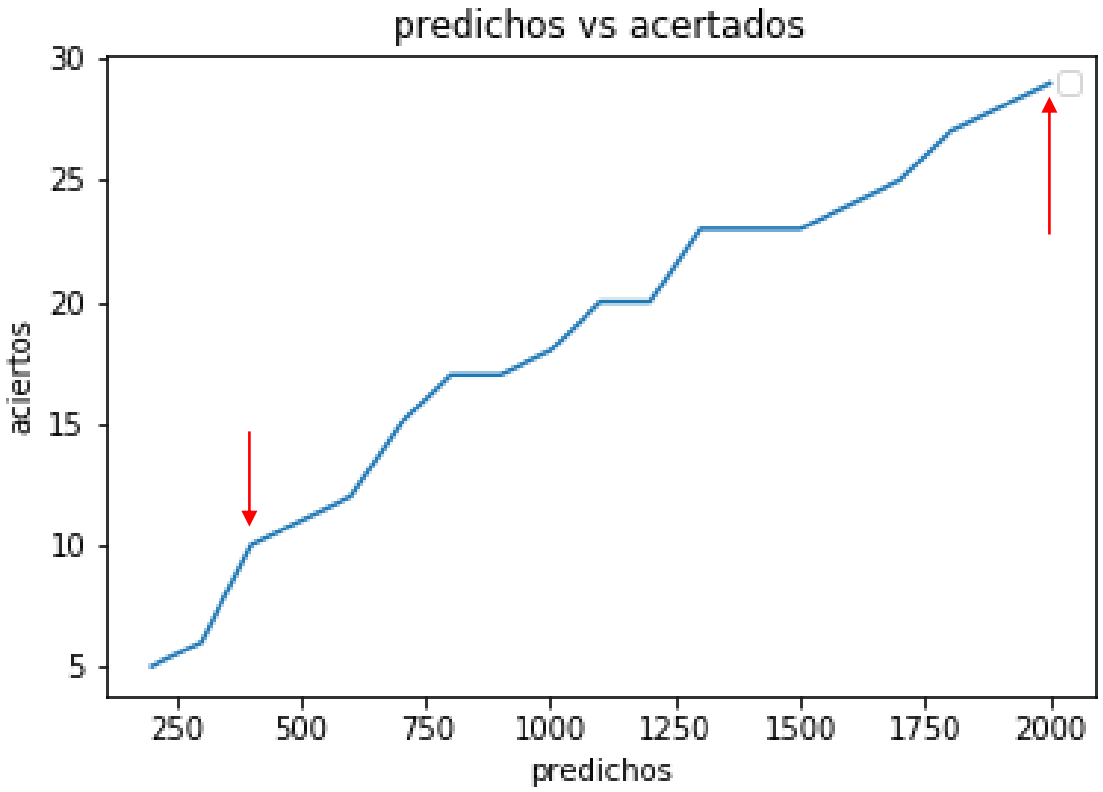


Tabla de resultados según puntos de corte

predicciones_positivas	aciertos	positivos_totales_a_predecir	porcentaje_aciertos	porcentaje_de_mejora
200	5	89	5.6%	0%
300	6	89	6.7%	1.1%
400	10	89	11.2%	4.5%
500	11	89	12.4%	1.1%
600	12	89	13.5%	1.1%
700	15	89	16.9%	3.4%
800	17	89	19.1%	2.2%
900	17	89	19.1%	2.2%
1000	18	89	20.2%	1.1%
1100	20	89	22.5%	2.2%
1200	20	89	22.5%	2.2%
1300	23	89	25.8%	3.4%
1400	23	89	25.8%	3.4%
1500	23	89	25.8%	3.4%
1600	24	89	27.0%	1.1%
1700	25	89	28.1%	1.1%
1800	27	89	30.3%	2.2%
1900	28	89	31.5%	1.1%
2000	29	89	32.6%	1.1%

Evaluación y Resultados del modelo mes de marzo 2020:

Resultados por azar:

Dotación convenio = 17600 personas
Accidentes marzo = 89
Proporción de accidentes =(89/17600)*100 = 0,5%

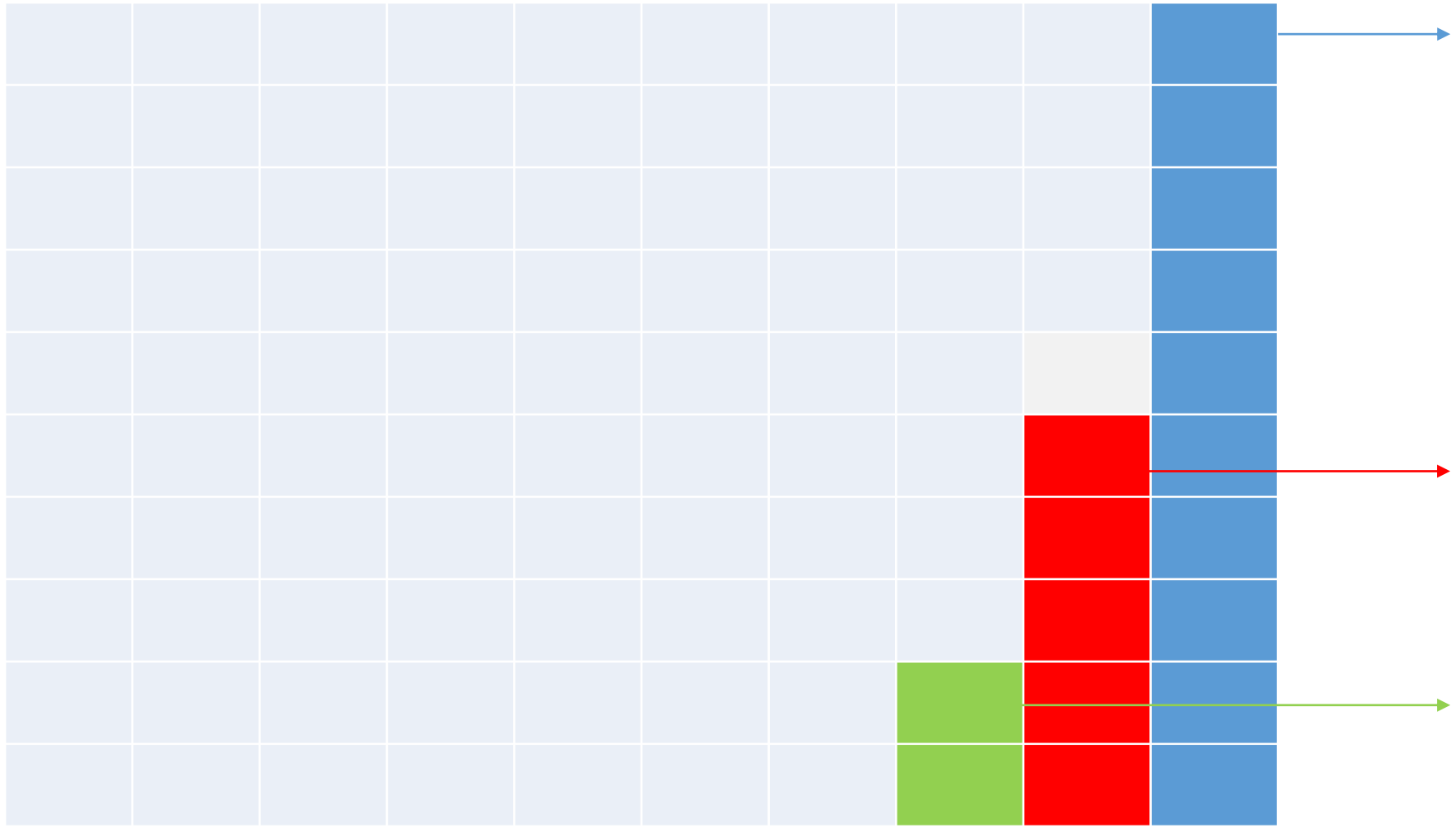
Si hubiesemos seleccionado 2000 personas **al azar** lo esperable hubiese sido detectar **10** siniestros. (0,5% * 2000 = **10**)

Resultados bajo el modelo:

Dotación convenio = 17600 personas
Accidentes marzo = 89
Proporción de accidentes =(89/17600)*100 = 0,5%

Al seleccionar 2000 personas **sugeridas por el modelo** se detectaron **29 siniestros.** Es decir, 3 veces mejor que el azar.

Resumen de resultados esperados en dotación dentro de convenio (17600 personas)



2000 predicciones pueden reducir entorno al **30%** de los accidentes **(360 anuales)**

1000 predicciones pueden reducir entorno al **20%** de los accidentes **(240 anuales)**

400 predicciones pueden reducir entorno al **10%** de los accidentes **(120 anuales)**

¿Por qué sabemos que funciona?

Datos conocidos para la construcción del modelo

Train set

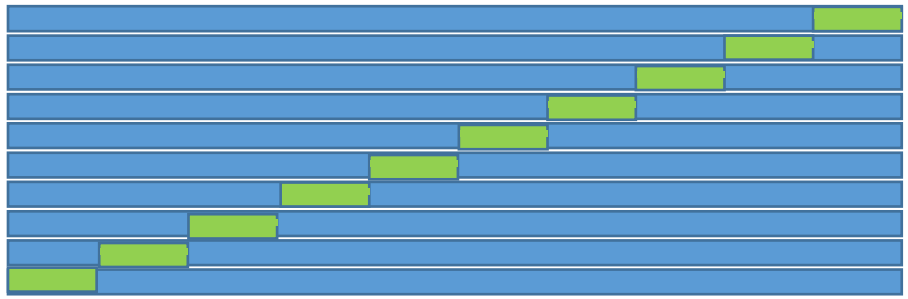
Validation set

Datos nunca vistos: Simulamos la realidad

Test set

Construimos el modelo con 1 año de historia. Por ejemplo: desde enero 2019 a enero 2020)

Por ejemplo: Con los datos de febrero 2020 predecimos marzo 2020



K – fold cross validation
Con K =10

Tabla Resumen de testeos en múltiples escenarios:

Condición	Mes predicción	Mejora respescto del azar	% Positivos detectados en 2000 predicciones
Estable_siniestros_completo	marzo_2020	entre 3 y 5 veces mejor	32,60%
Estable_siniestros_no_recurrentes	marzo_2020	entre 2 y 3 veces mejor	22,20%
Atípica_siniestros_completo	mayo_2020	entre 1,46 y 1,91 veces mejor	16,70%
Atípica_siniestros_no_recurrentes	mayo_2020	entre 1,4 y 1,85 veces mejor	16,10%

Conclusiones:

- Pese a haber sido evaluado en escenarios extremos, el modelo performa sistemáticamente mejor que el azar en todos ellos.
- En condiciones de estabilidad pruebas con 10-fold cross validation arrojaron un AUC de 74% (+- 2%). Para este tipo de problemas una puntuación de 70% se considera buena y una de 80% excelente.

¡Muchas Gracias!

People Analytics

CAPITAL HUMANO