Universidad Austral

Maestría en Ciencias de Datos

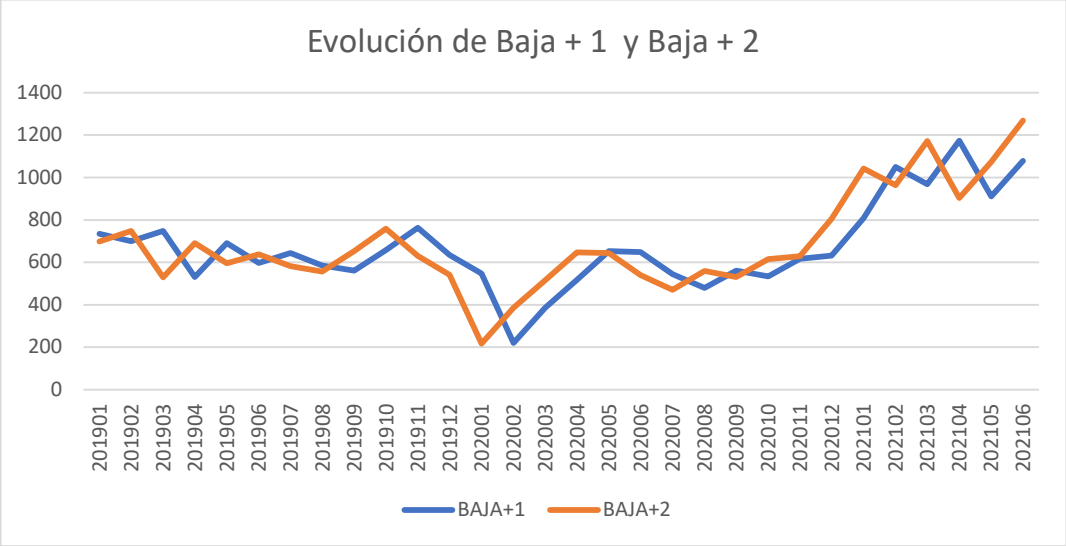
Laboratorio de Implementación I

Marcelo Giordano

Trabajo final:
Reducción de pérdidas por retención de clientes en entidad financiera

La secuencia de trabajo inició en los resultados de Experimentos Colaborativos, donde fue difícil encontrar conclusiones contundentes para aplicar en la configuración de scripts del workflow.

En el propio experimento realizado junto a Oscar Bianchi, que consistía en evaluar métodos de corrección de drift aplicados a las variables monetarias, ante la obtención de resultados desconcertantes dejamos planteado que era evidente que el cambio de comportamiento en los clientes a partir de 2021 condicionaba significativamente los resultados y para peor, siendo Julio 2021 el mes de futuro para el grupo B, cuando era el de mayor excepcionalidad de la serie.



Aquí cuesta encontrar el sentido a hacer mediciones de resultados con herramientas que no sirven para medir, más aún con el criterio ceteris paribus y limitarse a modificar únicamente el método de ajuste monetario.

En mi experiencia, el mismo inconveniente de pandemia debí resolverlo en la proyección de costos en prestaciones de salud de una prepaga de 250.000 asociados, para elaborar los presupuestos económico y financiero. Antes de cualquier medición económica, frente a la coyuntura de aislamiento se dividieron las prestaciones en impostergables (ej. parto, trat. Oncológico), postergables (ej. Cx de cadera), e irrecuperables (ej. Sesiones de psicología), de manera de proyectar sumando las impostergables y acumulando las postergables (más un plus por demora en tratamiento) para incrementar costos en la medida que se liberara la circulación,

que fue lo que terminó sucediendo: dos meses de baja demanda, luego creciente hasta superar el standard y finalmente se estabilizó en valores lógicos.

Con este ejemplo quiero referir que de alguna manera lo primero que debía realizarse antes de cualquier proyección era determinar la caída de bajas por efecto pandemia para luego normalizar el incremento una vez que la circulación volvió gracias a la aparición de las vacunas, que es lo que evidencia la curva del gráfico de bajas: Una caída abrupta en el inicio, una latencia por debajo de valores habituales hasta fines de 2020 y el rebote en 2021. Es fundamental ante cualquier proyección tener en claro cuando el fenómeno de crecimiento hará meseta y donde se espera que se normalice luego. Estamos hablando de comportamientos humanos, no es un modelo determinista de producción donde estimamos la posibilidad de productos defectuosos en el proceso de fabricación.

Vuelvo a referenciarme en mi experiencia. Para proyectar costos transformamos variables para expresarlos en una función lineal que salva la estacionalidad, la inflación y la "inflación médica", dado que la salud es la única industria donde la incorporación de tecnología no reemplaza a la anterior y agrega costos en lugar de reducirlos, el crecimiento de costos es un fenómeno global independiente de la inflación de cada país.

En el ejemplo se ve que a inicio de 2014 se produce una devaluación, estimamos que era "one shot", descartamos años de historia (la recta verde), proyectamos sobre los primeros meses de 2014 la performance del ejercicio Julio 14/Junio 15 (la recta naranja) y el comportamiento real fue el proyectado (la recta azul punteada).

Haber destinado todo este espacio a este tema es porque sinceramente creo que es crítico para resolver antes de cualquier aplicación de técnica (al menos con una propuesta razonable, a riesgo de errar).

Por lo tanto, la primera acción realizada para la competencia de Kaggle fue intentar validar las semi-pruebas de los experimentos colaborativos. Esto lo hicimos aplicando los métodos más relevantes sobre una extensión de tiempo con comportamiento más estable.

Y aquí comenzaron las dificultades: varias VM caídas, procesos interminables, etc. Por lo tanto decidimos cerrar el tema aplicando el método de normalización, que era el recomendable en base a unos pocos indicios (a pesar de haber dedicado mucho esfuerzo a resolver esto en los experimentos)

Inicio del Workflow:

711 CA Reparación del Dataset

De los experimentos ajenos, con algunas dudas similares al nuestro, consideramos tomar el método de "machine learning" para corregir los datos rotos.

721 DR Corregir Drifting

Lo ya expresado, tomé el método de normalización y no se realizan acciones sobre las variables intra-mes

731 FE historia

Lags y Tendencias no intento nada diferente a lo que traen los scripts, básicamente porque voy corriendo procesos con solo una variación entre sí para evaluar resultados (en lo que permite el espejismo de kaggle) y modificar estos parámetros hubiera generado ruido.

En una de las últimas corridas pongo en False Random Forest, no observo cambios significativos en los resultados y lo mismo con Canaritos Asesinos, lo fijo en 1.5 de ratio y 3.5 de desvío en base a experimentos de Julieta, Néstor y Sebastián. No observo variaciones significativas y dejo estos valores para lo siguiente, simplemente para no afectar la medición de otras pruebas, no por convicción de mejores resultados.

741 TS Training Strategy

Aquí vuelve a tomar peso el tema de los cambios de comportamiento en 2021. Del experimento de Jiang y Marcelo se verifica que correr hacia atrás, a inicios de 2021 el final del período de entrenamiento produce mejores resultados.

A pesar de no estar al inicio del WF, fue en donde comencé a realizar pruebas. Intento modificar los 3 meses no consecutivos de test ante la suposición que es el punto de corte más que el testeo lo que genera mejores resultados. Otra vez aquí se producen caídas de VMs, procesos interminables y demás contratiempos ya habituales. De alguna manera logro cerrar algunos procesos y ante resultados similares en el metro de goma que es kaggle decido dejar los 3 meses de testeo y no profundizar en este aspecto.

Finalmente, en las últimas corridas muevo el parámetro de undersampling de 0,4 a 0,5 simplemente para reducir tiempos ante la observación de no producir modificaciones significativas en los resultados e intentando correr más rápido otras pruebas. Claramente la secuencia lógica está invertida. Debiera correr todo el proceso con la mayor cantidad de datos y luego optimizar con canaritos y undersampling. Aquí fue al revés, con el solo objetivo de lograr un resultado aceptable de manera oportuna frente a uno de mejor calidad pero excedido del dead line de kaggle.

751 HT LightGBM

Cambio parámetros para Lambda 1, Lambda 2 y Main Gain To Split para que pasen por la bayesiana en base a valores estimados como razonables. Las iteraciones las mantengo en 50, estaba la intención de subirla, pero nuevamente el dead line me lleva a tomar esta decisión.

num_iteratio	seed	learning_rate	feature_fraction	num_leaves	lambda_l1	lambda_l2	min_gain_to_	estimulos	ganancia	iteracion_bay	rankeo
664	650011	0.0356163031524118	0.337420588841052	319	0	24	0.003393152	31693	145.707.566.716.642	29	1
3664	650011	0.0200047406592327	0.434170040449989	416	19	37	0.021132277	31737	145.296.766.116.942	78	2
328	650011	0.0361638183991635	0.562307197462961	557	0	20	5,64E+09	33752	144.237.613.193.403	31	3
3650	650011	0.0200186263116736	0.398451761416554	324	10	78	0.015423488	34384	144.203.682.158.921	55	4
5094	650011	0.0200444279256177	0.348481366250449	418	8	96	0.069233913	31859	144.025.698.650.675	62	5
1914	650011	0.0200894918481034	0.465958482443926	520	0	74	0.129960475	33592	143.827.658.170.915	46	6
4669	650011	0.0200455284996192	0.345669554379748	402	18	86	0.124387300	32459	143.784.019.490.255	71	7
3210	650011	0.0200219956936024	0.357954656556845	248	0	44	0.087130349	34142	143.642.185.907.046	67	8
1914	650011	0.0206540423764701	0.342944832485666	385	3	105	0.003073134	32690	143.582.773.613.193	36	9
3141	650011	0.0200218198848333	0.380049979039098	315	10	74	0.000338356	31906	143.508.206.896.552	59	10
1770	650011	0.0200210710342881	0.399202220486796	474	0	73	0.031044341	33558	143.505.370.314.843	54	11
872	650011	0.020328365567638	0.461278021774965	317	0	10	0.001187710	31887	143.430.514.242.879	43	12
1621	650011	0.0201062732104991	0.250985735885961	535	0	64	4,66E+09	29999	143.350.946.026.987	32	13
3796	650011	0.0200440313917117	0.353057216481711	329	6	76	0.072056216	32655	143.338.028.485.757	77	14
8475	650011	0.0200233653254313	0.336075482620435	383	13	85	0.111180472	30771	143.233.866.566.717	66	15
1446	650011	0.0201846378872739	0.38746061759399	402	0	34	3,46E+09	30183	143.141.854.572.714	35	16
3429	650011	0.0203856652237029	0.429640539239573	477	0	199	0.041490779	32835	143.130.802.098.951	40	17
1653	650011	0.0200192062353811	0.389365389512469	365	0	79	0.053839932	30821	143.118.289.355.322	53	18
4014	650011	0.0200203979851304	0.398119019133602	324	14	75	0.003615265	32497	143.116.856.071.964	56	19
4014	650011	0.0200203979851304	0.398119019133602	324	14	75	0.003615265	32497	143.116.856.071.964	57	20
6231	650011	0.0200026375102819	0.340361767425995	386	25	10	0.000188855	31284	143.056.965.517.241	37	21
3606	650011	0.0200922172840918	0.391825571200161	318	13	73	6,72E+09	34029	142.941.965.517.241	58	22
4473	650011	0.0200401424447722	0.388347395076088	528	0	66	0.318912259	35205	142.939.422.788.606	51	23
1697	650011	0.0204154278544929	0.44055866355328	374	0	64	2,04E+09	26970	142.860.254.872.564	41	24
2410	650011	0.0200421806497848	0.350077585711872	327	0	83	0.093759482	30081	142.841.557.721.139	61	25
2410	650011	0.0200421806497848	0.350077585711872	327	0	83	0.093759482	30081	142.841.557.721.139	75	26
1993	650011	0.020016408255644	0.336595479982827	411	0	96	0.075856109	32135	142.770.352.323.838	64	27
3415	650011	0.0202745844942695	0.366193206948113	425	0	27	0.411482299	31359	142.687.317.841.079	38	28
4590	650011	0.0200570200715814	0.45895264311368	556	0	79	0.198657665	30895	142.279.197.901.049	47	29
2006	650011	0.0206044012768071	0.32878572223891	29	0	0	0.081163297	32288	142.199.028.485.757	30	30
3940	650011	0.0200009966862619	0.441266206717504	503	14	74	0.007987305	35805	142.123.650.674.663	48	31
5040	650011	0.0203609771122578	0.501866163755489	10	19	31	0.001828336	28870	142.075.097.451.274	60	32
3254	650011	0.0201523823526476	0.349644431405363	584	0	147	0.039396690	35155	141.953.410.794.603	42	33
8611	650011	0.0200017933589618	0.345505435411802	359	15	69	0.034838636	31108	141.898.308.845.577	74	34
3934	650011	0.020007707900493	0.384343287591036	511	0	65	0.319256818	28638	141.796.164.917.541	52	35
1774	650011	0.0200233076644794	0.451913570693034	399	0	64	0.012564305	34960	141.648.575.712.144	49	36
2968	650011	0.0200474710411108	0.343385464757733	289	0	83	0.050101303	28532	141.580.011.994.003	68	37
4556	650011	0.0200486601095262	0.459848371708154	94	0	135	0.148593097	27528	141.575.460.269.865	45	38
2458	650011	0.0783690624193051	0.228238013723692	488	2	174	0.668436820	32420	141.388.155.922.039	18	39
374	650011	0.0498030501963064	0.360535715092395	722	0	8	0.030719651	33557	141.323.362.818.591	34	40
3254	650011	0.0201545586075369	0.331618002919409	703	1	127	0.878412344	30035	141.309.092.953.523	76	41
3957	650011	0.0341323563362309	0.644526060869142	968	73	198	0.338930385	32293	138.387.986.506.747	27	42
393	650011	0.141548188298946	0.316029046984981	543	61	70	0.230621452	35281	138.145.905.547.226	13	43
584	650011	0.0432226142327651	0.366982360364335	439	0	0	0.003354137	28399	137.538.292.353.823	44	44
138	650011	0.267357722632293	0.808307501051566	170	25	12	0.027421187	33601	137.527.254.872.564	19	45

En la optimización bayesiana elijo iteraciones que presentan buenos resultados con algunos parámetros particulares, eligiendo los rankings 1, 3, 12, 22, 30, 31 y 45 para semillerío.

771 ZZ Semillerío.

De los rankings anteriores logro correr para 20 semillas a 1, 3 y 12. 22 lo descarto y 31 luego de 3 semillas, testeo resultados en kaggle y no ofrece nada brillante por lo que lo aborto y corro para 10 semillas a los ranks 30 y 45 para llegar sobre la hora de cierre de la competencia.

781 ZZ Hibridación.

Hago diferentes pruebas, teniendo al rank 1 como de mejor performance por sí solo. Ante las restricciones de tiros en kaggle, en varios ranks testeo entre 10.000 y 11.000 envíos, ante la observación que son los envíos habitualmente de mejor performance.

	NF r1 s20	**NF r3 20**	**NF r12 s20**	**NF r30 s10**	**NF r45 s10**
9500	53,53903	50,29909		51,36907	52,45905
10000	54,06902	51,72906	50,90907	52,16905	51,69906
10500	54,10902	52,15905	50,66908	51,70906	52,27905
11000	53,06904	50,65908	52,70904	52,16905	52,79904
11500	52,72904			51,58906	51,78906
Prom.	53,50303	51,21157	51,42906	51,80106	52,20505

Realizo distintas hibridaciones probando varias combinaciones. Los valores no muestran variaciones significativas, no obstante, lo ya mencionado sobre la verificación contra kaggle que no es una referencia de medición y en consecuencia no hay certeza de nada de lo que se está observando.

	r1+r3+r12	**r1+r3**	**r1+r12**	**r3+r12**	**r1+r3+r12+r30+r45**		
	Único	**unico**	**unico**	**unico**	**Simple**	**Pond**	
9500	52,53905	52,4391	52,4391		52,07905	52,4991	52,399050
10000	52,37905	52,859	52,939	50,9091	52,49905	52,4191	52,334050
10500	52,89904	52,899	53,259	51,8091	52,22905	52,2091	52,550712
11000	52,29905	52,629	53,049	50,7091	52,65904	52,2291	52,262383
11500	51,54906	51,5891	51,8991		51,67906	51,5891	51,661060

Ante la incertidumbre, opto por la hibridación con mayor cantidad de elementos y el nivel de envíos que en promedio mejores resultados arrojó en las diferentes hibridaciones.

Conclusiones:

Resultó interesante y con grandes expectativas emprender este trabajo. Claramente el tiempo disponible es escaso para un abordaje acorde a lo que implica dentro de la maestría la asignatura.

Transitar el workflow permite distinguir etapas de manera muy clara, diversos procesos y parámetros involucrados, aunque solo desde un punto operativo y agregaría, de bajo nivel. En las expectativas había dos grandes focos: algoritmos (programación) e interpretación de resultados que retroalimentan esos algoritmos.

Prácticamente no hubo desarrollo de programación, solo intervenir en scripts ajenos y en una librería de R (data table) no utilizada hasta el momento en la maestría, lo que implica un gasto de energía en meras cuestiones de sintaxis que de manera infalible una vez que logran resolverse y correr, exponen cruelmente que en el fragor de la batalla, reduciendo sentencias a su mínima expresión para resolver el problema, la lógica de formulación ha terminado destrozada, en un bochorno absoluto.

Por el lado de interpretación de resultados no fue mejor la cosa. Lo ya dicho del espejismo de kaggle, midiendo resultados con una herramienta mentirosa poco aporta a interpretar algo. Y los tiempos de proceso suman a la causa para tener pocas oportunidades de prueba para jugar y al menos por las consecuencias de una acción inferir lo que sucede. Las acciones terminan siendo intuitivas, por azar o por lo que fuera, pero claramente lejos de lo científico. Quizás para esto hay una capilla en la universidad.

Todo esto en un marco donde los tiempos de dedicación excedieron ampliamente a los planteados como propuesta de la maestría. Hasta un punto llevé la cuenta de horas, cuando pasó el cuádruple de lo declarado dejé de hacerlo por lo innecesario que era a esa altura. No solo en cantidad, también en calidad, teniendo una notebook a mano permanentemente para cada dos o tres horas revisar, corregir, rehacer, etc. etc. Sumando el desatino de la universidad que nuevamente ajusta su cronograma disponiendo de los recesos, que es donde generalmente las personas programan sus vacaciones familiares, acorde a la lógica con que acuerda tácitamente la sociedad. Haber programado esta materia, con lo que implica, con el dedicado desarrollo que compromete el docente y con la entrega del trabajo final en el receso, creo que es una falta de respeto. Habría que recordarle a la universidad que hasta el mismísimo Señor al 7mo. día descansó.

Adicionalmente a esto, sumando las horas destinadas, las horas de dedicación que aportó el profesor Denicolay, párrafo aparte para su disposición y omnipresencia 7x24 ante cada minúscula consulta, el costo de Google, que por el hecho de no pagarse no deja de pesar en mi caso aprox. 450 dólares, creo que el despilfarro de recursos para comprimir en 8 semanas lo que debiera insumir como mínimo un semestre termina de dejando sensación a poco en relación a lo que bueno que aportó frente a lo que insumió. En síntesis, percibo el mismo provecho que obtendría un estudiante de arquitectura descargando un camión de ladrillos.