Experimento #5

Agregado de variables con Random Forest

Equipo A:
Pablo Marcelo Amado
Tomás Correa

Hipótesis Experimental del Problema #05

- **Hipótesis A** :El agregado de variables al dataset generadas de forma automática por un modelo de Random Forest mejora las ganancias obtenidas por el modelo.

- **Hipótesis B:** Al menos 1 variable queda como una de las 10 más importantes del modelo.

Bibliografía

Si bien no hemos encontrado una bibliografía extensa específicamente sobre la experimentación con random forests, nos hemos apoyado en los siguientes artículos para guiar nuestra investigación:

- Amat Rodrigo, J. (2017). Árboles de decisión, random forest, gradient boosting y C5.0. Ciencia de Datos. https://cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting
- Adele, G. (2021, 10 de septiembre). Using random forests to generate partially synthetic, categorical data. Medium. https://medium.com/@gideonadele/using-random-forests-to-generate-partially-synthetic-categorical-data-4d2b6a664988

Sesgos Cognitivos (conflictos de intereses)

- Nuestras expectativas de que la implementación de variables devenidas del random forest tenga un impacto significativo en las ganancias podrían introducir un sesgo en la interpretación de los resultados.

- La convicción de que la optimización automatizada de los hiperparámetros del random forest era clave para maximizar las ganancias de manera eficiente nos llevó a demorar el diseño nuestro experimento.

Diseño experimental

- En primer lugar, realizamos una ejecución del script z509_workflow_base sin agregar variables generadas por el modelo Random Forest, con el objetivo de obtener una curva de ganancia para comprar. A esta ejecución la llamamos Control

- Luego se realizan varias ejecuciones del mismo script, modificando los hiper parámetros del modelo random forest. Estos son: num_iterations, num_leaves, num_data_in_leaf, feature_fraction_bynode. Las ejecuciones son las siguientes:

Diseño experimental

Hiperparámetros	Experimento 01	Experimento 02
num_iterations	20	50
num_leaves	16	40
num_data_in_leaf	1000	200
feature_fraction_bynode	0.2	0.1

Diseño experimental

- Por último, comparamos las curvas de ganancia promedio de las tres ejecuciones para comprobar si la hipótesis A es verdadera o no.

- En todas las ejecuciones, comprobamos si existen variables generadas por el modelo random forest en el top 10 de la lista de importancia de variables del modelo final.

Limitaciones

El presente estudio se enfrentó a diversas limitaciones que podrían haber influido en los resultados obtenidos:

- **Experiencia limitada en investigación:** La falta de experiencia previa en este tipo de investigación pudo haber afectado la capacidad de anticipar y abordar ciertos desafíos metodológicos.
- **Conocimientos de programación limitados:** La poca experiencia en programación dificulta la comprensión de los scripts utilizados, lo que ralentizó el proceso de experimentación y pudo haber limitado la capacidad de realizar modificaciones o adaptaciones necesarias.
- **Restricciones en los recursos computacionales:** La dependencia de máquinas virtuales spot, debido a la falta de recursos, generó problemas en algunas ejecuciones y en la carga de resultados a Kaggle, lo que pudo haber afectado la eficiencia y la reproducibilidad del experimento.
- **Limitaciones temporales:** El tiempo disponible para realizar el experimento fue limitado, lo que impidió explorar soluciones más extremas o realizar un análisis más exhaustivo de los resultados.

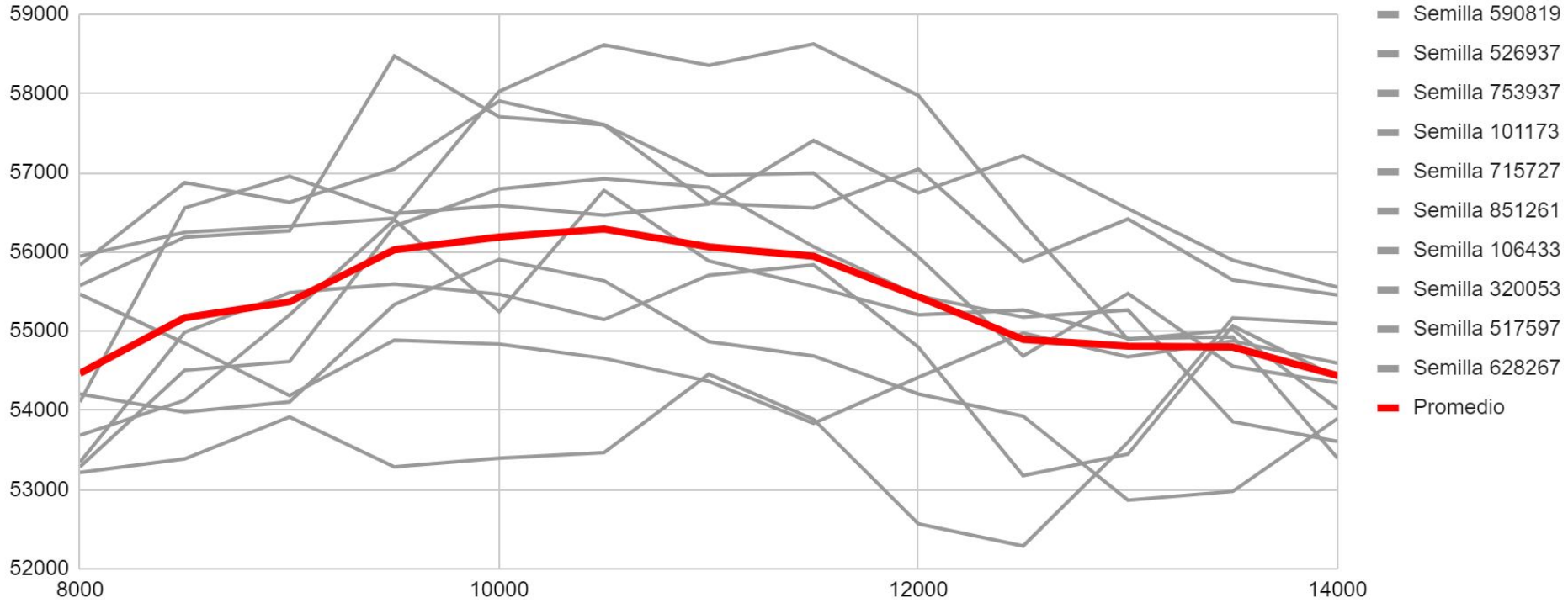
Resultados

Resultados de ejecución control (Sin variables RF):

Envíos	Semilla 590819	Semilla 526937	Semilla 753937	Semilla 101173	Semilla 715727	Semilla 851261	Semilla 106433	Semilla 320053	Semilla 517597	Semilla 628267	Promedio
8000	55577	53287	55947	54107	53347	55467	54207	53217	53687	55837	54468
8500	56187	54507	56247	56557	54987	54847	53977	53387	54127	56877	55170
9000	56267	54617	56327	56957	55487	54187	54107	53917	55207	56627	55370
9500	58477	56327	56427	56487	55597	54887	55337	53287	56407	57047	56028
10000	57707	56797	58027	56587	55467	54837	55907	53397	55247	57907	56188
10500	57607	56927	58617	56467	55147	54657	55637	53467	56777	57607	56291
11000	56617	56817	58357	56607	55707	54367	54867	54457	55887	56967	56065
11500	56557	56067	58627	57407	55837	53837	54687	53887	55567	56997	55947
12000	57047	55447	57977	56747	54797	54417	54207	52567	55207	55937	55435
12500	55877	55177	56357	57217	53177	54977	53927	52287	55267	54687	54895
13000	56417	55267	54897	56547	53447	54677	52867	53597	54907	55477	54810
13500	55647	53857	55017	55897	55067	54877	52977	55167	54927	54557	54799
14000	55457	53607	54017	55557	54417	54597	53897	55097	53397	54347	54439

Resultados

Resultados de ejecución control (Sin variables RF):



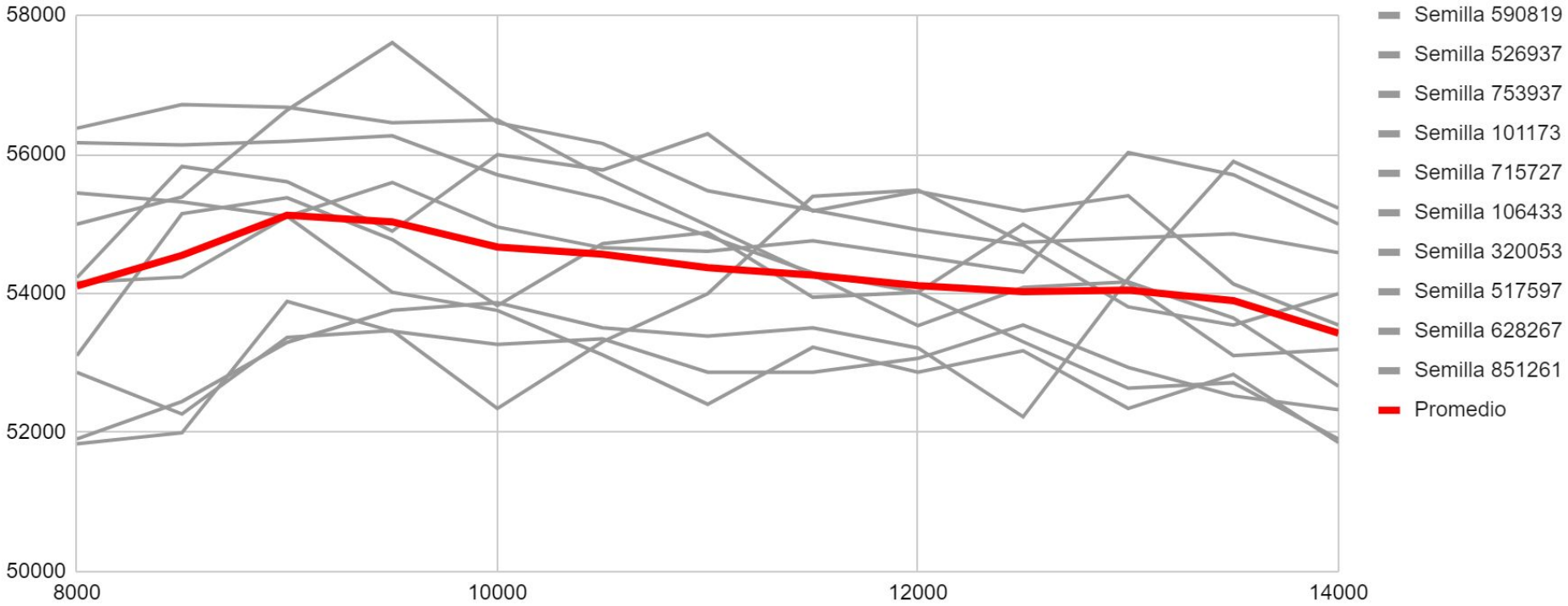
Resultados

Resultados de ejecución Experimento 01 (Workflow Base):

Envíos	Semilla 590819	Semilla 526937	Semilla 753937	Semilla 101173	Semilla 715727	Semilla 106433	Semilla 320053	Semilla 517597	Semilla 628267	Semilla 851261	Promedio
8000	56377	56167	55447	54997	54227	51907	53107	54157	52867	51837	54109
8500	56717	56137	55317	55387	55827	52447	55147	54237	52267	51997	54548
9000	56677	56187	55107	56637	55607	53297	55377	55107	53367	53887	55125
9500	56457	56267	55597	57607	54897	53757	54777	54017	53467	53457	55030
10000	56497	55707	54957	56457	55997	53867	53827	53757	52347	53267	54668
10500	55687	55367	54657	56157	55777	53507	54717	53117	53307	53347	54564
11000	54977	54827	54607	55477	56297	53387	54877	52407	53997	52867	54372
11500	54267	54297	54757	55197	55187	53507	53947	53227	55397	52867	54265
12000	53537	54017	54537	54917	55467	53217	54017	52867	55487	53067	54113
12500	54087	54997	54307	54697	55187	52227	53307	53177	54737	53547	54027
13000	54167	54147	56027	53807	55407	54217	52637	52347	54797	52937	54049
13500	53647	53107	55707	53547	54137	55897	52717	52837	54857	52527	53898
14000	52667	53197	54997	53997	53547	55227	51907	51857	54587	52327	53431

Resultados

Resultados de ejecución Experimento 01(Workflow Base):



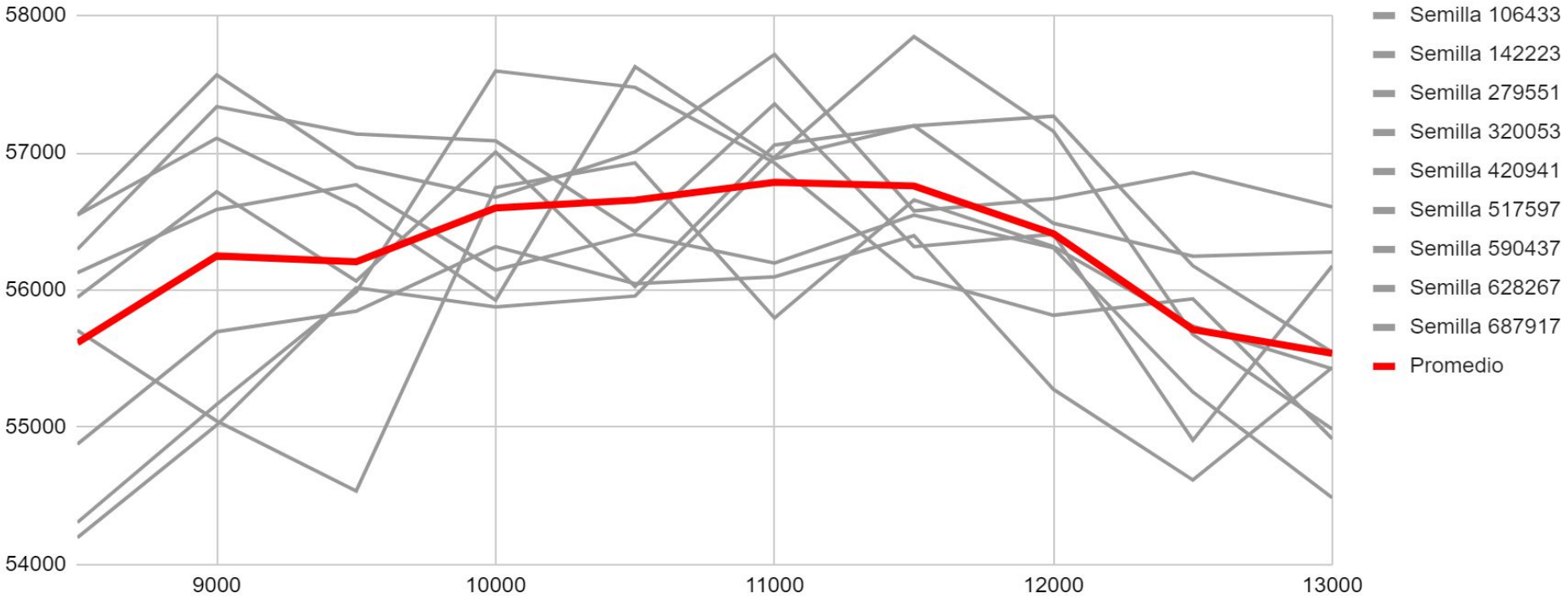
Resultados

Resultados de ejecución Experimento 02:

Envíos	Semilla 106433	Semilla 142223	Semilla 279551	Semilla 320053	Semilla 420941	Semilla 517597	Semilla 590437	Semilla 628267	Semilla 687917	Promedio
8500	54877	56547	55707	54197	54307	56127	56547	55947	56297	55617
9000	55697	57107	55047	55017	55167	56587	57567	56717	57337	56249
9500	55847	56607	54537	56017	55987	56767	56897	56067	57137	56207
10000	56317	55927	56747	55877	57597	56147	56677	57007	57087	56598
10500	56047	57627	56927	55957	57477	56407	57007	56027	56427	56656
11000	56097	56957	55797	56967	56927	56197	57717	57057	57357	56786
11500	56397	57197	56657	57847	56097	56547	56577	57197	56317	56759
12000	55277	56487	56317	57157	55817	56307	56667	57267	56407	56411
12500	54617	56247	55737	55677	55937	55257	56857	56177	54907	55713
13000	55437	56277	55427	54987	54917	54487	56607	55547	56177	55540

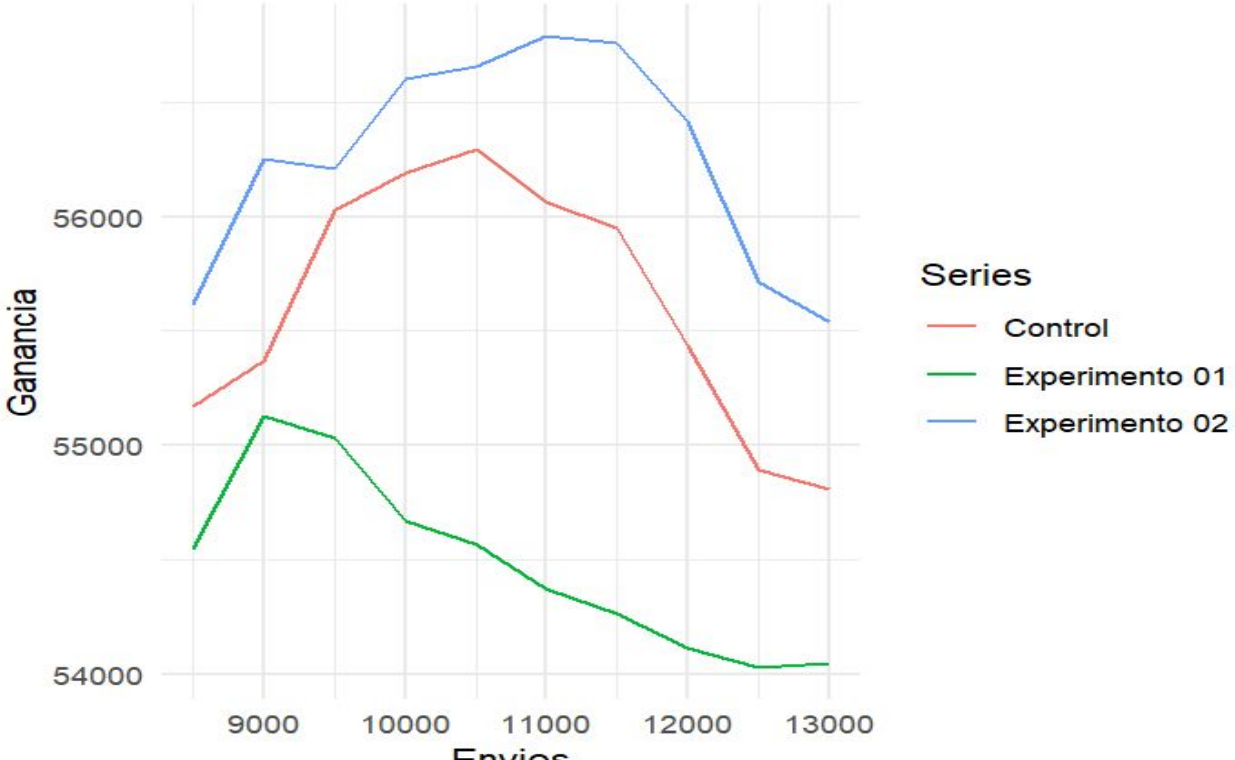
Resultados

Resultados de ejecución Experimento 02:



Resultados

Comparación de curvas de ganancia:



Aplicando el test de Wilcox a los vectores de ganancia promedio del experimento 02 y el control, se obtiene un p-valor de 0.0115. Con eso comprobamos que las ganancias del experimento 02 son mayores al control.

Resultados

Importancia de las variables en los 3 experimentos:

Ranking	Control		Experimento 01		Experimento 02	
	Feature	Gain	Feature	Gain	Feature	Gain
1	ctrx_quarter_normalizado	0,068709899	**rf_015_004**	0,07689309	**rf_014_025**	0,095635906
2	ctrx_quarter_normalizado_lag1	0,039705331	ctrx_quarter	0,040839365	ctrx_quarter_normalizado	0,078048198
3	ctrx_quarter	0,037687697	cpayroll_trx	0,03876346	mcuentas_saldo_rank	0,02439392
4	cpayroll_trx	0,030798109	**rf_001_004**	0,031097113	mprestamos_personales_rank	0,024175907
5	mcaja_ahorro_rank	0,019837553	ctrx_quarter_normalizado	0,030613115	ctrx_quarter	0,022652939
6	mprestamos_personales_rank	0,016102906	**rf_013_004**	0,02969744	ctrx_quarter_normalizado_lag1	0,013670515
7	mcuentas_saldo_rank	0,01556099	mpayroll_sobre_edad_rank	0,022408342	cpayroll_trx	0,012621061
8	vm_msaldototal_rank	0,014484839	mcaja_ahorro_rank	0,02161038	**rf_046_005**	0,012105221
9	ccomisiones_mantenimiento_tend6	0,014281621	ctrx_quarter_normalizado_lag1	0,021012473	mcaja_ahorro_rank	0,011196876
10	mpayroll_sobre_edad_rank	0,010786648	**rf_020_006**	0,019726671	cdescubierto_preacordado_tend6	0,008437874

Cantidad de variables Generadas:

- Experimento 01: 260
- Experimento 02: 1615

Discusión (de los resultados)

- Observamos que el experimento 02 es el que mayor ganancia obtiene, superando levemente a la ejecución control.
- El experimento 01 (Workflow Base) presenta ganancias considerablemente más bajas que el experimento 02 y a la ejecución control.
- En el top 10 de importancia de variables, encontramos variables generadas por el Random Forest. En el experimento 01 hay 4 variables y en el experimento 02 hay dos. Debemos señalar que en ambos casos la variable de mayor importancia fue generada por el modelo de Random Forest.
- Se puede ver en los gráficos de cada experimento, que la ejecución control presenta mayor variabilidad en la ganancia.

Conclusiones

- El agregado de variables aleatorias a través de un modelo de Random Forest mejora las ganancias del modelo predictivo.
- A medida que se aumenta la cantidad de variables, la ganancia mejora. Para lograr esto, se deben realizar mayores iteraciones, con árboles más profundos y utilizando pocos datos en cada hoja.
- Cuanto más variabilidad en los nodos, los árboles son bien diferentes entre sí, lo que mejora la ganancia del modelo final.
- En todos los modelos generados por Random Forest, se crean variables de importancia para el modelo final, encontrando al menos una variable en el top 5 de importancia.

Recomendación concreta

- Generar Variables procedentes de Random Forest
- Utilizar los siguientes hiper-parámetros:
  - num_iterations: 50 - 60
  - num_leaves: 40 - 50
  - num_data_in_leaf: 100 - 200
  - feature_fraction_bynode: 0.1
- Considerar la utilización de canarios asesinos para reducir el tiempo de cómputo

Futuros Problemas y Experimentos

- Cómo próximos experimentos consideramos que sería interesante realizar la ejecución del modelo aplicando dos o más veces el generador de variables de random forest, de manera que la segunda ejecución genere variables considerando las generadas anteriormente.
- Otro posible experimento interesante sería probar una optimización de hiperparametros de random forest teniendo en cuenta las limitaciones temporales, ya que esto conlleva un gran gasto de cómputo para realizar.
- Por último, se podría probar modificar los meses en los que se entrenan los modelos de random forest, y comprobar si entrenar en una mayor cantidad de meses genera mejores variables.

Anexo

Repositorio Github:

1. https://github.com/tomascorrea93/ITBA_MINERIA_Exp5_GrupoA_RF_