Reporte Capstone Project

Jairo Ordóñez Guardado

Contexto

El banco PortuBank ha venido invirtiendo mucho dinero en campañas de telemercadeo para colocar sus productos de captación de recursos ofreciendo a sus clientes contratos de depósitos de dinero a plazo. Últimamente se ha visto que a pesar de contactar a muchos clientes, no se está logrando colocar cuantos contratos se quisieran y los costos por las campañas de telemarketing se han ido incrementando de manera significativa.

Esta situación tiene preocupados a los altos jerarcas quienes han solicitado se analice la situación para ver de qué forma pueden ser más eficientes en las campañas de telemercadeo. En este punto es que contactaron a nuestra firma, AnaliTicos para efectuar un estudio del caso y determinar si es posible ser más precisos al momento de contactar a los clientes y tener una mayor probabilidad de aciertos en la colocación de los depósitos a plazo.

Se ha logrado obtener una data con información de clientes, campañas y otros datos donde se conoce si un cliente aceptó o no efectuar un depósito a plazo. El objetivo es utilizar estos datos para desarrollar un modelo que pueda responder a la siguiente pregunta **¿El cliente aceptará SÍ o NO realizar un depósito a plazo con nuestra entidad bancaria?** De lograr este objetivo, las campañas de telemarketing de PortuBank podrían ser más intencionales con aquellos clientes cuya información suponga una mayor probabilidad de acierto según el análisis que se desarrolle.

Objetivos

Nos hemos planteado los siguientes objetivos:

- Abordar la data y prepararla para las distintas tareas que requiere el análisis.
- Analizar los factores principales relacionados a que un cliente acepte o no un contrato de depósito con el banco.
- Evaluar diferentes modelos que nos permitan predecir la respuesta del cliente.

Algunas de las hipótesis que nos hemos planteado al darle un vistazo general a los datos fueron las siguientes:

- Creemos que los clientes con mejores posiciones laborales pueden estar más abiertos a la posibilidad de efectuar depósitos a plazo con el banco.
- De la misma manera el nivel educativo puede estar asociado a una mayor aceptación de contratos de depósito, sobre todo aquellos clientes con mayor preparación.
- En cuanto a rango de edades, consideramos que de los 35 a 50 años puede ser el grupo que mejor respuesta positiva tenga hacia los depósitos, considerando que podría ser una población más estable laboralmente.
- Respecto a los datos de la campaña de telemercadeo nos interesa saber como incide el método de contacto así como la cantidad de llamadas, los días y meses cuando se efectúan las mismas.

Tratamiento de Datos

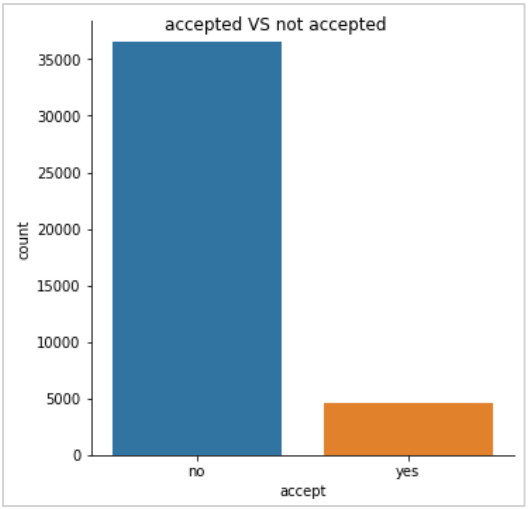
Para abordar este análisis implementamos algunos pasos y técnicas comunes en machine learning, entre ellos:

- Tareas de preparación de datos como: renombrar variables, convertir tipos de variables, discretización de datos.

- Ingeniería de características: a partir de variables existentes creamos nuevas variables usando la técnica de label encoding para tratar las variables categóricas.
- Selección de características: para hacer un poco más eficientes los modelos implementamos la técnica de Recursive Feature Elimination y así obtener un subset de datos con una selección basada en un método modelado y no una escogencia al azar.
- Resúmenes de datos mediante tablas pivot así como distintas visualizaciones en función de los datos que se querían analizar.
- Modelado de datos con diferentes tipos de clasificadores donde incluimos:
    - División de datos en training y testing.
    - Cross Validation.
    - Entrenamiento y predicción de los modelos.
    - Reporte de métricas principales.

Visualizaciones y Hallazgos

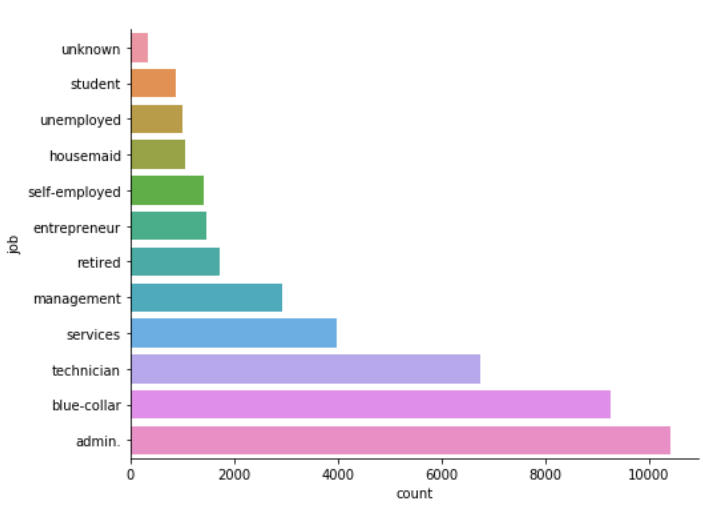
Al iniciar este análisis lo primero que nos interesaba hacer era tener un vistazo de la distribución de los contratos de depósitos aceptados vs los no aceptados para darnos una idea general del panorama. Lo que encontramos es lo que vemos en la imagen al lado.

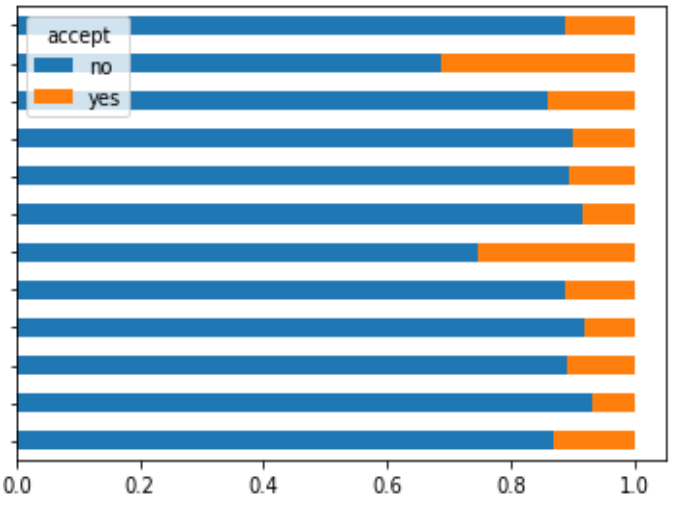


Como se puede ver, la distribución de respuesta está mucho más inclinada hacia NO aceptar los planes de depósito, esto a simple vista puede parecer muy negativo para la entidad bancaria puesto que solo el 11% de los clientes dicho SÍ a los depósitos y el restante 89% ha declinado.

Para determinar que tan bueno o malo puede ser este porcentaje de aceptación tendríamos que hacer benchmarking y comparar la eficiencia del banco en campañas de telemercadeo con otras entidades que ofrezcan planes de depósito similares, pero este alcance está fuera de los límites del análisis actual.

En cuanto al **status laboral**, nuestra percepción inicial era que los clientes con mejores trabajos pueden tener mayores probabilidades de aceptar depósitos a plazo. A la izquierda el número de clientes y a la derecha el % de aceptación.

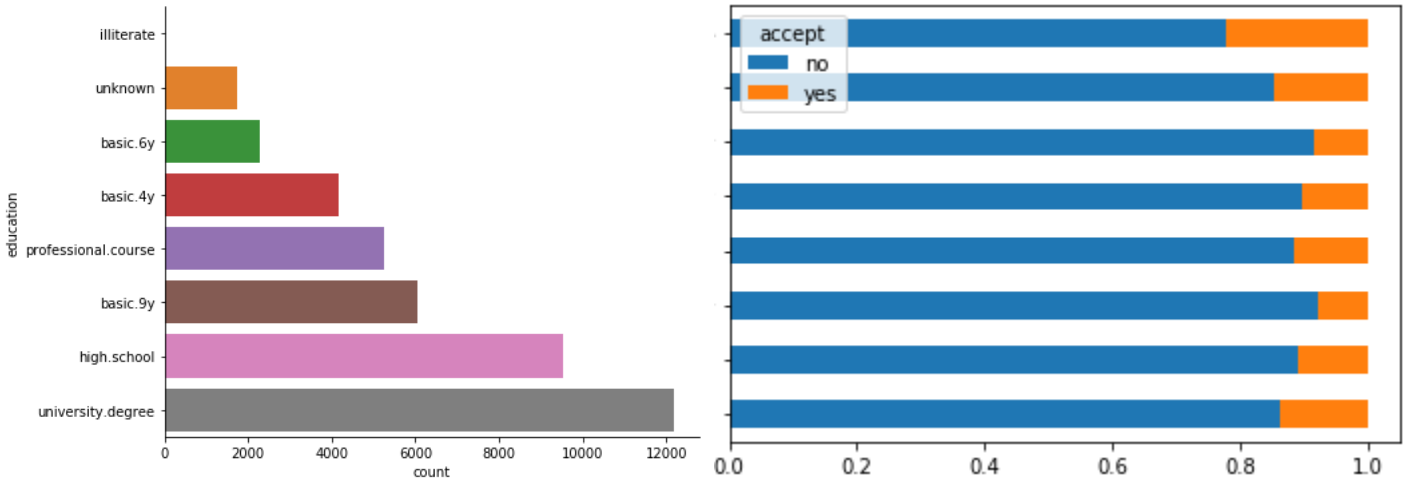




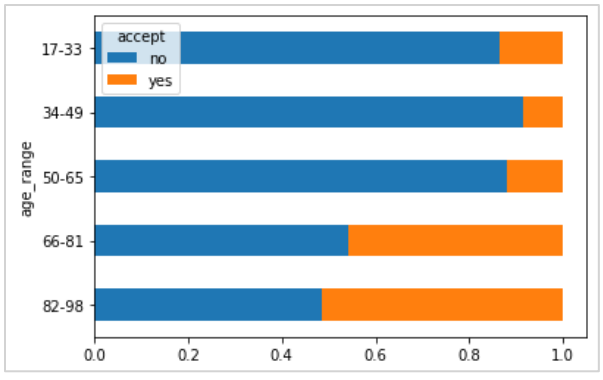
Lo interesante de las gráficas anteriores es que las posiciones laborales que más aceptan depósitos con el banco son los Estudiantes (31%) y los Retirados (25%). Por otro lado los clientes que más se contactan en estas campañas de telemercadeo son los Técnicos, Operarios (blue-collar) y Administrativos. Sin embargo, estos no muestran porcentajes de aceptación tan altos como los mencionados antes (11%, 7% y 13% respectivamente).

Si hay más efectividad tratando con Estudiantes y Retirados, ¿por qué no hacer que el telemercadeo del banco tome mayor fuerza sobre esta población de clientes? Al final y bajo este panoráma que muestran los datos, se pudo constatar que el nivel laboral no está ligado a mejores resultados en la aceptación de depósitos con el banco.

Otra aspecto que queríamos observar era el comportamiento de los depósitos aceptados según el **nivel educativo** de los clientes. Acá también nuestra espectativa era encontrar una relación que a mejor nivel educativo mejor proporción de aceptación, las gráficas siguientes nos muestran lo encontrado:



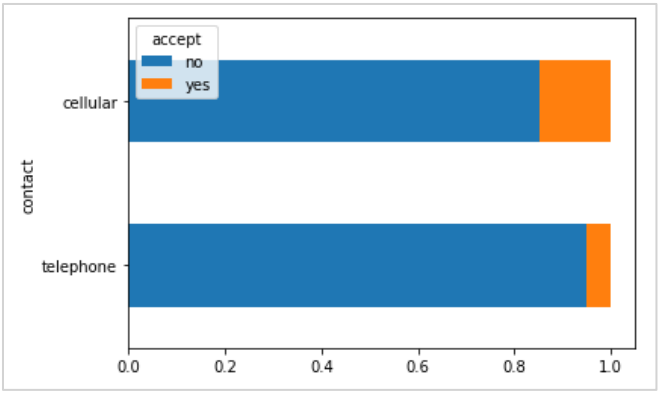
Si bien es cierto nuestra hipótesis no está del todo refutada, pues los clientes con grado universitario están entre los que mayor porcentaje de aceptación muestraron (14%), algo que si llama la atención es como los clientes catalogados como “iliterados” son los que mayor aceptación tienen con un 22%, el problema es que es un grupo con muy pocos clientes. Sería interesante profundizar más en este grupo, que más podríamos saber sobre esta categoría de clientes, cómo podríamos alcanzar más de ellos.



Respecto a la **edad de los clientes** pensábamos que el rango de los 35 a los 50 años sería el grupo que mejor aceptación a depósitos con el banco presentaría, esto creyendo que es un rango de edad donde las personas en teoría están más estables laboralmente y ya han tenido la oportunidad de desarrollar una carrera laboral. No obstante, contrario a nuestra hipótesis, este rango de edad tiene la proporción más baja, tan solo el 9% de los clientes en este rango de edad acepta un depósito con el banco. Los datos nos mostraron que los clientes sobre los 65 años son los que más aceptan contratos de depósitos bancarios, con un porcentaje de hasta el 50%. Esto tiene mucha relación con lo que anteriormente presentamos respecto al status laboral de los clientes, ya que en ese apartado también aparecían los “Retirados” entre las categorías que mejor respuesta afirmativa tenían hacía los depósitos.

Lo que también es un hecho, es que el grueso de los clientes llega hasta el límite de los 50 años; sin embargo, la efectividad de las campañas de telemercadeo hasta ese límite no es tan alto como en personas de mayor edad. El reto aquí es ver como aumentar las tasas de aceptación en estos grupos de edad donde más clientes se contactan y por otro lado aumentar el número de clientes en edades mayores donde mejor proporción de aceptación tenemos.

Respecto a los datos de la campaña de telemercadeo nos interesaba revisar algunas variables que consideramos importantes y que podían darnos vistas estratégicas de la efectividad que se ha venido teniendo. Los hallazgos más importantes que encontramos los presentamos a continuación.

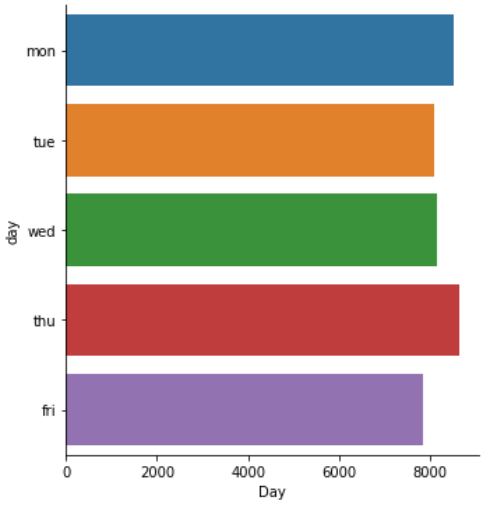


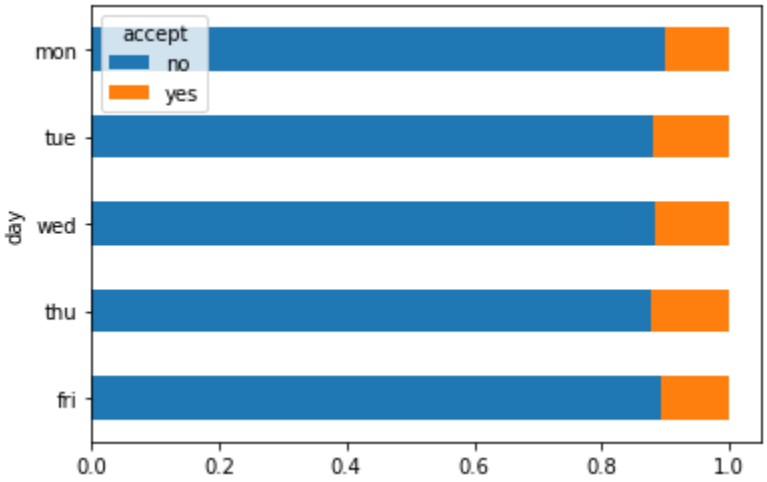
Primero quisimos observar los **medios de contacto** utilizados en la campaña, a los clientes se les ha contactado de dos maneras: teléfono fijo y celular. Lo que se pudo constatar es que el grueso de llamadas se lleva a cabo a celulares tal como nos muestra la tabla al lado y de la misma manera es mediante esta vía como se obtiene una mayor proporción de aceptación por parte de los clientes. De las personas contactadas a sus celulares, el 15% aceptó un plan de depósito a plazo mientras que los que fueron contactados a teléfono fijo solo el 5%.

contact	contact_
cellular	26144
telephone	15044
All	41188

Evidentemente se tiene una mejor respuesta cuando se contacta a los clientes a sus celulares.

Otro factor que visualizamos fueron los **días en los que se contactaron a los clientes**. Si vemos el gráfico debajo de la izquierda vemos la cantidad de contactos por día que se efectuaron, los cuales no presentan variaciones tan significativas. En el gráfico a la derecha tenemos la proporción de aceptación de depósitos según cada día de la semana.

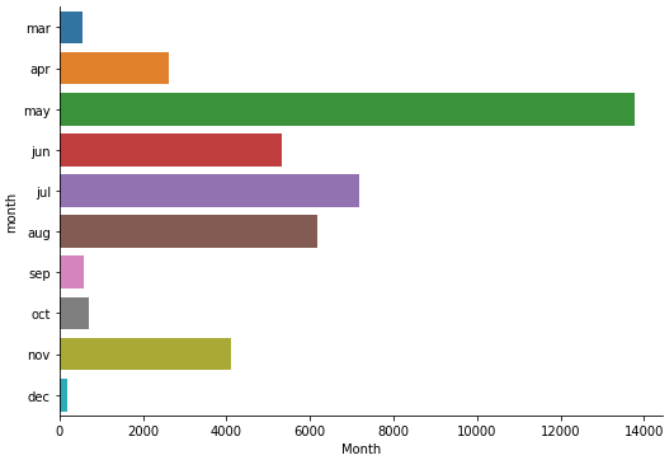


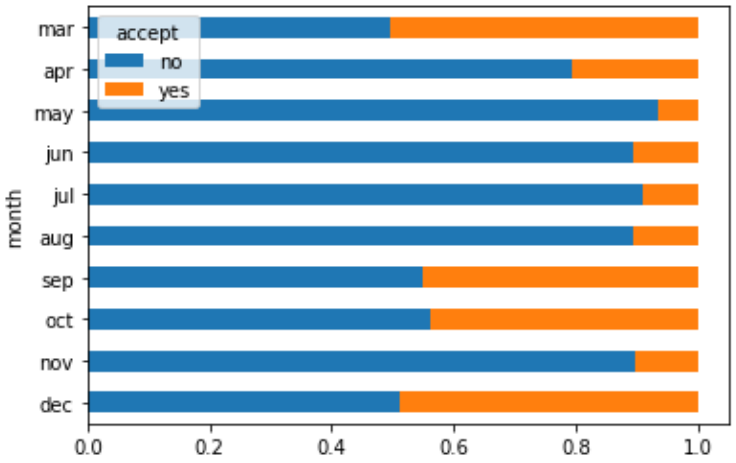


Durante la actual campaña de marketing los lunes y viernes se muestran como los días cuando menos contratos de depósitos aceptados se lograron (10% y 11% respectivamente). Mientras que de martes a jueves levemente la respuesta es más positiva (12%).

¿Estará esto relacionado a un comportamiento social? Pensemos en el lunes como el primer día de trabajo de la semana y quizá los clientes no estén de ánimo para una llamada de esta naturaleza, o en el caso de los viernes, el último día de la semana y los clientes tengan su mente en los planes del fin de semana. Parece que lo más estratégico es abordar a los clientes en los días medios de la semana.

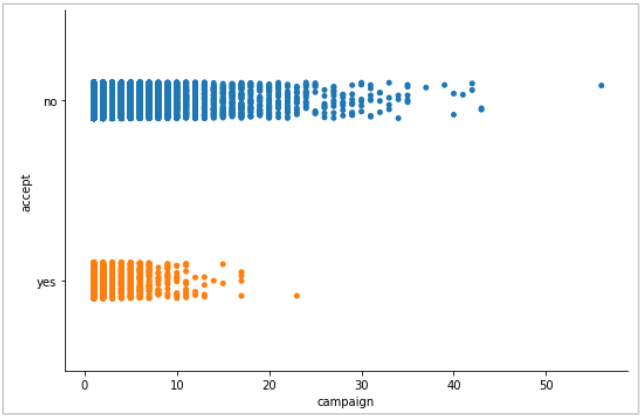
Este mismo ejercicio lo hicimos con los datos correspondientes a los **meses** y nos encontramos con hallazgos interesantes.





Los meses donde menos llamadas se efectuaron son los que mejor proporción de depósitos aceptados tuvieron: marzo (51%), diciembre (49%), setiembre (45%) y octubre (44%). Mientras que en mayo cuando el telemercadeo fue más intenso, los resultados positivos fueron bajos (6%). En enero y febrero no se realizaron contactos a clientes.

Sería interesante profundizar en las razones del por qué durante esos cuatro meses cuando más contratos fueron aceptados la campaña de marketing no fue tan intensa como en otros meses. ¿Qué otros factores de contexto que no están en los datos pueden ayudar a averiguar esto y sobre todo llevar al banco a concentrar sus esfuerzos de marketing en esos meses cuando mejor respuesta positiva se obtiene?

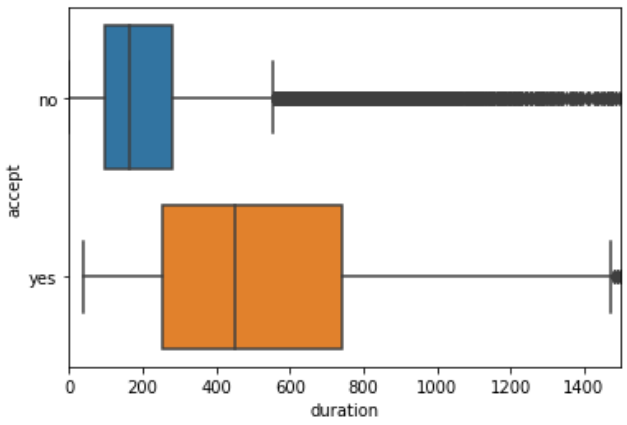


La **cantidad de llamadas** relacionada con la aceptación o no de un depósito con el banco muestra un dato interesante.

La gráfica al lado nos dice que los clientes que respondieron SÍ al contrato de depósito recibieron menos llamados que los que se negaron. Esto refleja la insistencia por parte de los encargados de las llamadas hacia los clientes, sin embargo también podría dar una sugerencia de cuantos intentos podrían ser “suficientes” para tener una respuesta positiva.

Viendo el gráfico, este número de intentos podría tener un límite de 15 llamadas por cliente.

El último dato que nos interesaba observar era el de la **duración de las llamadas**. La siguiente gráfica solo nos confirma algo que es fácil de suponer: las llamadas donde los clientes SÍ aceptaron un contrato de depósito con el banco registraron duraciones mayores respecto a los clientes que NO aceptaron la propuesta. En la gráfica hay una cantidad significativa de valores atípicos que no se observan por el zoom efectuado en la escala, los cuales podrían estar asociados a la cantidad de observaciones que presenta el dataset. Esto se suma a que las distribuciones de los datos no son normales.



El dato de la duración de la llamada no se conoce hasta que la llamada se realiza, por lo tanto este dato no se contemplará en los modelos más adelante. No obstante, hay un alcance que consideramos importante a partir de este dato: si observamos la gráfica, se nota que el rango de duración donde la mayoría de clientes aceptaron un contrato podría estar entre los 200 y 800 segundos (de 3 a 13 minutos), esto podría convertirse en un factor a considerar cuando las llamadas estén en curso y alertar a un

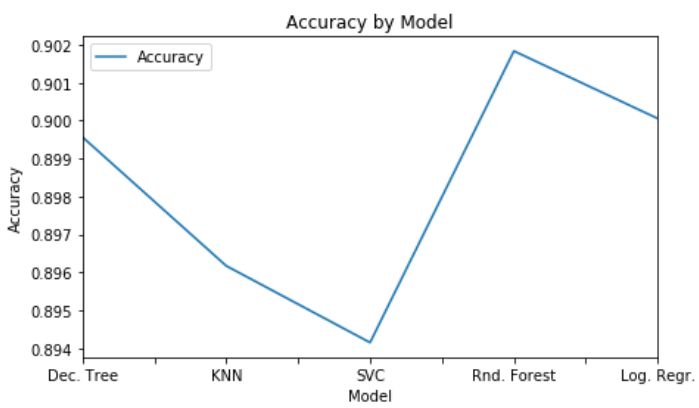
colaborador de telemercadeo que está ante un cliente que potencialmente puede aceptar la oferta del depósito a plazo, por lo tanto el colaborador debe tratar de no perder ese chance.

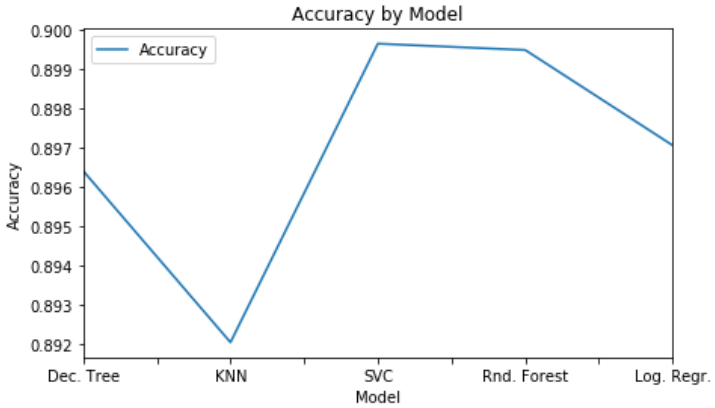
Modelos

En el apartado de modelado llevamos a cabo dos experimentos, en ambos casos omitimos la variable correspondiente a la duración de la llamada pues este dato es desconocido antes de efectuar contactos a los clientes. En el primer experimento trabajamos con todas las variables disponibles para observar que resultados obteníamos de esta forma, mientras que para el segundo experimento nos inclinamos por una selección de características sugerida mediante la técnica de Recursive Feature Elimination.

En ambos experimentos dividimos los datos en una proporción 70/30 (training/testing) y utilizamos validación cruzada a 10 folds. Además llevamos a cabo predicciones con cada modelo, así como sus matrices de confusión y reportes de métricas principales.

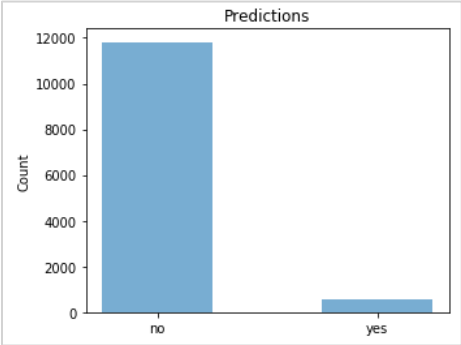
Al estar frente a un problema de clasificación, todos los algoritmos utilizados fueron para este tipo de casos. Probamos cinco diferentes clasificadores en cada uno de los dos experimentos. Las métricas de exactitud que obtuvimos en cada uno de los experimentos las resumimos a continuación. A la izquierda las métricas del modelo con todas las características y a la derecha con selección usando RFE.





Como se puede ver, ambos experimentos muestran métricas de exactitud bastante buenas (90% en promedio). A pesar de ser muy similares las diez opciones posibles, nuestra recomendación se inclinaría por utilizar el clasificador Random Forest bajo una selección de variables RFE. Consideramos que este algoritmo es bastante robusto pues para sus estimaciones involucra un conjunto de árboles de decisión, además al usar un subset de datos definido mediante RFE, estamos asegurando que se utilizaran las variables que mayor peso tienen en la respuesta que el modelo necesita brindarnos y por ende también bajamos en cierta medida el costo computacional del modelo en la fase de entrenamiento.

Con el Random Forest + RFE obtuvimos los siguientes resultados:



	precision	recall	f1-score	support
no	0.91	0.98	0.95	10973
yes	0.63	0.25	0.36	1384
accuracy			0.90	12357
macro avg	0.77	0.61	0.65	12357
weighted avg	0.88	0.90	0.88	12357

Conclusiones y Recomendaciones

Con mencionamos al inicio de los hallazgos, la proporción positiva de contratos de depósitos colocada en clientes en apariencia es muy baja, apenas de un 11% contra un rechazo del 89%. Esto de primera mano hace preguntarse ¿vale la pena este tipo de campañas para la captación de fondos?

Necesitaríamos tener más información que determine si ese 11% es realmente malo o está dentro de lo aceptable, para esto podría ser de apoyo algún tipo de benchmarking que nos permita conocer cuál es la tasa de colocación de productos similares de otras entidades financieras para compararlos con los que obtiene el banco en estudio.

Además, otro dato que podría haber agregado valor el estudio es una variable que indicara el monto de depósito contratado en aquellos casos donde el cliente aceptó lo oferta, esto por lo menos nos daría una idea de cuánto dinero se ha podido captar bajo esta modalidad y así conocer si son sumas importantes para el banco.

Y si queremos ir a niveles más macros sería interesante observar que proporción del total de ingresos captados por parte del banco corresponden a depósitos obtenidos mediante campañas de marketing y conocer si su participación es significativa. Todo esto podría dar una idea más clara de que tan rentable o no resulta ser ese 11% de aceptación que este tipo de campañas está generando.

Con lo que se pudo analizar hay algunos aspectos que podemos puntualizar:

- Si a nivel de tipo de trabajo, los Estudiantes y Retirados son los que más aceptan hacer depósitos de capital en el banco, ¿por qué no hacer que las campañas de telemercadeo traten de llegar a más clientes con este perfil laboral?
- Algo similar sucede a nivel educativo, los clientes catalogados como iliterados son los que mejor respuesta positiva muestran. Consideramos importante conocer más acerca de estos clientes, se podrían encontrar datos que justifiquen tratar de contactar a más clientes de este tipo. No podríamos desechar esto hasta no profundizar un poco más, a simple vista podríamos pensar que este grupo no tendría mucha liquidez que ofrecer, pero esto es algo que solo ahondando más se podría confirmar.
- A nivel de rangos de edades, parece que hay un nicho muy importante sobre los 65 años el cual corresponde a clientes ya retirados. Este grupo de clientes muestran la proporción más importante respecto a hacer depósitos a plazo con el banco (sobre el 50%). Se podrían hacer esfuerzos por tratar de contactar a más clientes en este rango de edad.
- A nivel de días en los que se contactan a los clientes, los datos muestran que se tiene un poco más de efectividad cuando las llamadas se hacen de martes a jueves. Esto lo podríamos entender como un factor social, lunes y viernes son días donde el sentimiento de los clientes puede estar en otras direcciones.
- Con respecto a los meses cuando se efectúan los contactos a los clientes, fue posible observar divergencias significativas: en los meses donde mejor respuesta afirmativa se tiene es cuando menos se contacta a los clientes. Podría ser de ayuda y agregando más variables del tipo fecha, ver si a nivel de series de tiempo es posible profundizar el tema y determinar tendencias o factores de estacionalidad que permitan llevar a cabo estas campañas en periodos más efectivos.