Universidad Cenfotec
Data Analytics
Módulo 5-Capstone Project
Ana Vanessa López Monge

BANK MARKETING ANALYSIS

1. Información General

Contexto: Dentro de la industria bancaria, optimizar la focalización para el telemercadeo es un tema clave debido a una presión creciente para aumentar las ganancias y reducir los costos. La crisis financiera de 2008 cambió drásticamente el negocio de los bancos europeos. En particular, los bancos portugueses fueron presionados para aumentar los requisitos de capital, por ejemplo, al capturar más depósitos a largo plazo.

Objetivo principal: Diseñar e implementar un sistema de soporte de decisiones (DSS) efectivo e inteligente que utilice un enfoque de minería de datos (DM) para la selección estratégica de clientes de telemercadeo bancario.

Beneficios:

-Repensar las estrategias de mercadeo enfocándose en maximizar el valor del cliente a través de la evaluación de la información disponible y las métricas del cliente, lo que permite construir relaciones más largas y más estrictas en alineación con la demanda comercial.

-Seleccionar el mejor conjunto de clientes, es decir, que aquellos que por sus características es más probable que suscriban del producto.

Conjunto de datos: El conjunto de datos presenta 41188 observaciones y 21 variables.

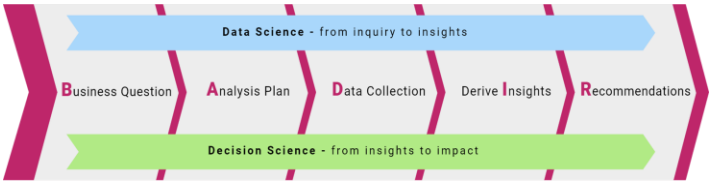
Variable dependiente: Variable discreta y binaria “Deposito a plazo”.

Variables independientes: Se cuenta con un total de 21 diferentes variables continuas y discretas. Dentro de estas variables se cuenta con atributos de telemercadeo, detalles del producto e información del cliente.

Hipótesis principal: ¿Se puede predecir si el cliente aceptará el Depósito a largo plazo dependiendo de las variables de telemercadeo, detalles del producto e información del cliente?

2. Metodología

Para este proyecto utilizamos la metodología BADIR, el cual es un marco de referencia muy efectivo para el manejo de proyectos relacionados con minería de datos y transformación digital.



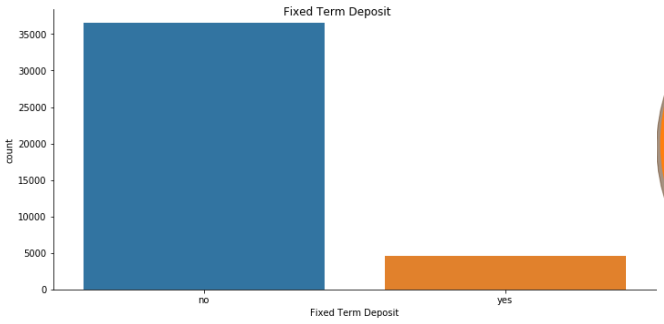
Universidad Cenfotec
Data Analytics
Módulo 5-Capstone Project
Ana Vanessa López Monge

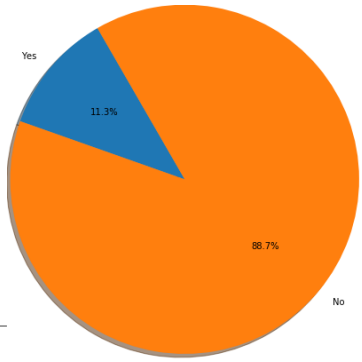
3. Procedimiento

4. Hallazgos y Resultados

4.1. Analisis Explortorio de la Data

- Por medio del análisis de correlación, covarianza y mapa de color, se puede observar que no hay una relación lineal fuerte entre la variable dependiente (Deposito a plazo) y las diferentes variables dependientes. Estas herramientas sí nos permiten observar cierta colinearidad, es decir existe una correlación fuerte entre algunas variables predictoras.
- Se puede observar una tasa baja de aceptación de los depositos a plazo, 11% vs una tasa de rechazo de 89%. Lo anterior justifica la necesidad de optimizar la focalización en torno a estrategias de marketing más efectivas.



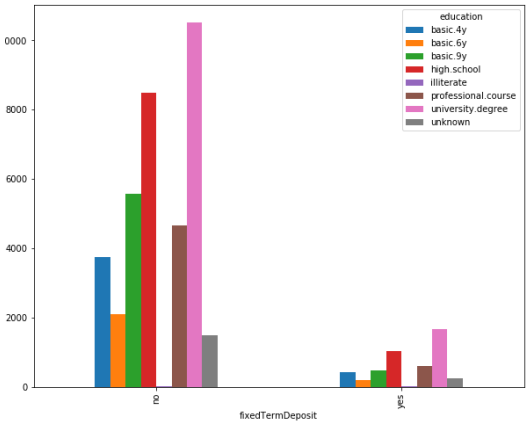


- La mayoría de los clientes se contactan por medio de celular. Además proporcionalmente un 15% de las personas que se contactan por celular aceptan el deposito a plazo mientras que un 5% lo hacen por teléfono.

fixedTermDeposit	no	yes
contact		
cellular	0.852624	0.147376
telephone	0.947687	0.052313

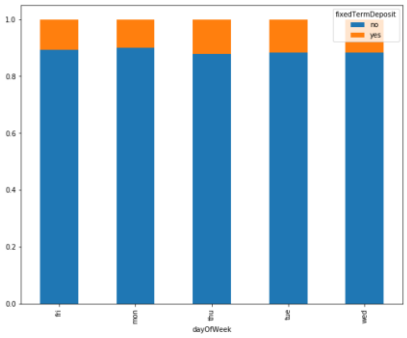
Universidad Cenfotec
Data Analytics
Módulo 5-Capstone Project
Ana Vanessa López Monge

- Gráficamente se observa una diferencia significativa en la tasa de aceptación del deposito dependiendo del nivel educativo. Aquellos clientes con grados de university y unknown son los que muestran proporcionalmente mayor aceptación a los depositos a plazo (14% en promedio).



fixedTermDeposit	no	yes
education		
basic.4y	0.897510	0.102490
basic.6y	0.917976	0.082024
basic.9y	0.921754	0.078246
high.school	0.891645	0.108355
illiterate	0.777778	0.222222
professional.course	0.886515	0.113485
university.degree	0.862755	0.137245
unknown	0.854997	0.145003

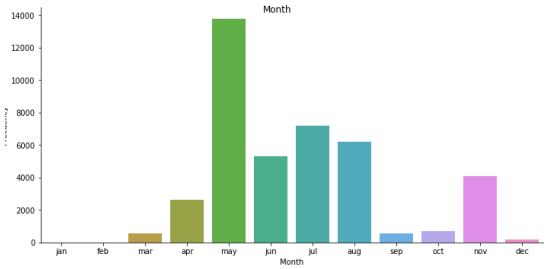
- Gráficamente no se observa una diferencia significativa en la tasa de aceptación del deposito dependiendo del día de la semana en que se contactan a las personas. En promedio la aceptación de depositos a plazo diaria es de un 11%. Complementariamente a los gráficos, la prueba de hipótesis chi realizada nos muestra que con nivel de confianza del 95% hay una dependencia significativa entre la variable día y la variable a predecir.



dayOfWeek	no	yes	All
fri	19.10	18.23	19.00
mon	20.98	18.25	20.67
thu	20.73	22.52	20.94
tue	19.53	20.54	19.64
wed	19.66	20.45	19.75
All	100.00	100.00	100.00,

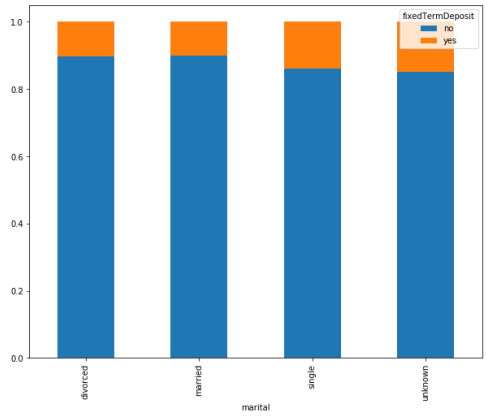
	Chi-square test	results
0	Pearson Chi-square (4.0) =	26.1449
1	p-value =	0.0000
2	Cramer's V =	0.0252)

- El mes donde se realizan más llamadas es mayo, sin embargo este mes es el que muestra una de las tasas de aceptación más bajas con un 6%. Proporcionalmente los meses de marzo, septiembre, octubre y diciembre son los que proporcionalmente muestran mejores tasas de aceptación por lo que sería importante aumentar el volumen de llamadas durante estos meses.



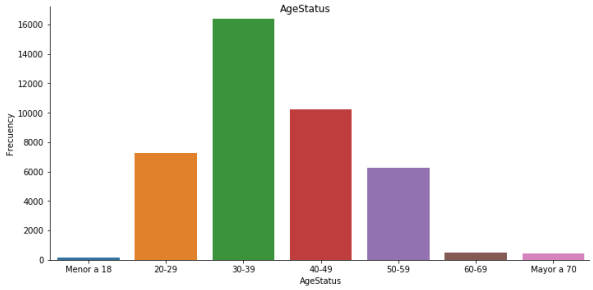
Universidad Cenfotec
Data Analytics
Módulo 5-Capstone Project
Ana Vanessa López Monge

- Gráficamente se visualiza que las personas casadas son a las que se contactan con mayor frecuencia, sin embargo los clientes solteros y estatus desconocido son los que proporcionalmente mostraron mayor tasa de aceptación de los depositos a plazo comparado con los casados, 14% y 15% vs 10% respectivamente. Complementariamente a los gráficos, la prueba de hipótesis chi nos muestra con un nivel de confianza del 95% que hay una dependencia significativa entre la variable estatus civil y la variable a predecir.



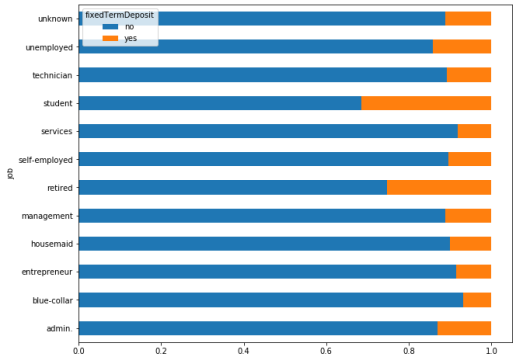
	no	yes	All
marital			
divorced	11.32	10.26	11.20
married	61.28	54.57	60.52
single	27.22	34.91	28.09
unknown	0.19	0.26	0.19
All	100.00	100.00	100.00,
	Chi-square test		results
0	Pearson Chi-square (3.0) =		122.6552
1	p-value =		0.0000
2	Cramer's V =		0.0546)

- Visualmente se observa que el banco se ha enfocado en contactar mayoritariamente a las personas en el rango de 30-39, sin embargo proporcionalmente los clientes mayores a 70 y entre 60-69 son los que muestran mejores tasas de aceptación, 47% y 43% respectivamente.



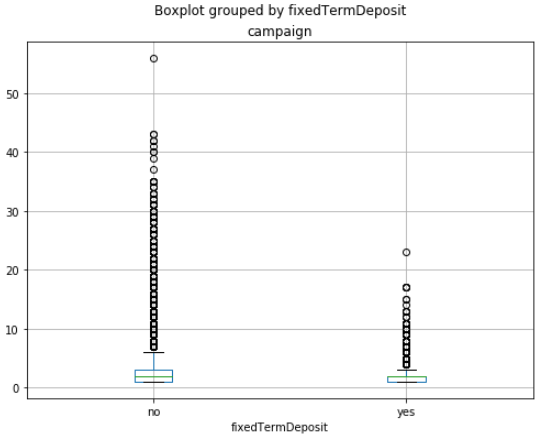
fixedTermDeposit	no	yes
ageRange		
Menor a 18	0.592857	0.407143
20-29	0.852685	0.147315
30-39	0.902533	0.097467
40-49	0.918262	0.081738
50-59	0.893461	0.106539
60-69	0.565574	0.434426
Mayor a 70	0.521327	0.478673

- A pesar de que los clientes bajo un perfil de estudiante y retirado no están entre los más contactados, proporcionalmente sus tasas de aceptación son de las más altas con un 31% y 25% respectivamente.



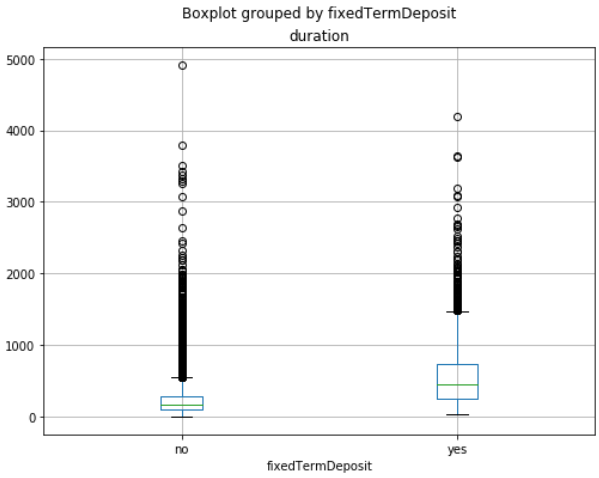
Universidad Cenfotec
Data Analytics
Módulo 5-Capstone Project
Ana Vanessa López Monge

- El boxplot de # Número de veces que una persona fue contactada vs la variable dependiente, muestra una diferencia significativa en las distribuciones de los datos y se puede visualizar que aquellos que aceptaron el deposito a plazo fueron contactados menos veces en comparación con los que lo rechazaron. Complementariamente se realiza una ANOVA de un factor y se concluye con un 95% de confianza que las medias en el # de contactos de ambos conjuntos (aceptación vs rechazo) difieren entre sí.



	df	sum_sq	mean_sq	F	PR(>F)
fixedTermDeposit	1.0	1391.562959	1391.562959	182.156673	2.007780e-41
Residual	41186.0	314635.259513	7.639374	NaN	NaN

- El boxplot de Duración de la llamada vs la variable dependiente Depositos a plazo, muestra una diferencia significativa en las distribuciones de los datos y se puede visualizar que aquellos que aceptaron el deposito a plazo registraron duraciones de llamadas mayores con respecto a los que rechazaron el deposito. Complementariamente se realiza una ANOVA de un factor y se concluye con un 95% de confianza que las medias en el # de contactos de ambos conjuntos (aceptación vs rechazo) difieren entre sí.



	df	sum_sq	mean_sq	F	PR(>F)
fixedTermDeposit	1.0	4.547710e+08	4.547710e+08	8094.101634	0.0
Residual	41186.0	2.314055e+09	5.618548e+04	NaN	NaN

4.2. Modelos Predictivos

Se realizaron específicamente dos experimentos para resolver el problema de clasificación para predecir si el cliente aceptará el deposito a plazo o no. Para ambos experimentos se aplicaron diferentes algoritmos de clasificación tales como Random Forest, Knn, Super Vector Machine y árbol de decisiones.

En ambos experimentos dividimos los datos en una proporción 70/30 (entrenamiento/testeo) 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.

- **Experimento 1**

En este experimento se toman en cuenta todas aquellas variables/atributos disponibles en el conjunto de datos y se realiza un trabajo de ingeniería de variables en aquellas variables independientes de tipo categóricas.

A continuación se muestran las métricas de desempeño resumen obtenidas para los diferentes algoritmos:

	Performance metrics per model-Experiment 1					
	Training	Testing				
	Accuracy	Precision No	Precision Yes	Recall No	Recall Yes	Accuracy
Decision Tree	0,94	0,94	0,61	0,95	0,55	0,91
Super Vector Machine	0,90	0,91	0,65	0,98	0,22	0,91
Knn	0,92	0,93	0,63	0,96	0,49	0,90
Random Forest	0,92	0,92	0,74	0,99	0,31	0,91

Tanto durante la fase de entrenamiento y como de testing se obtienen métricas muy similares lo que nos permite garantizar la existencia de consistencia en los modelos durante ambas fases. Las métricas de los algoritmos son muy similares, sin embargo Random Forest muestra mejores resultados en cuanto a precision y recall. Las predicciones obtenidas con Random Forest se detallan a continuación:

Prediction True	0	1	All
0	10758	160	10918
1	987	452	1439
All	11745	612	12357

- **Experimento 2**

Se toman únicamente aquellas variables/atributos que sugiere la función RFE (Recursive Feature Elimination). Las 10 variables independientes con un efecto más significativo sobre la variable a predecir segun RFE son: ['age','duration','campaign', 'pdays','consPriceIdx', 'euribor3m', 'nrEmployed', 'job1', 'education1', 'dayOfWeek1'].

A continuación se muestran las métricas de desempeño resumen obtenidas para los diferentes algoritmos:

	Performance metrics per model-Experiment 2					
	Training	Testing				
	Accuracy	Precision No	Precision Yes	Recall No	Recall Yes	Accuracy
Decision Tree	0,94	0,94	0,61	0,96	0,53	0,91
Super Vector Machine	0,90	0,91	0,62	0,98	0,20	0,91
Knn	0,92	0,94	0,63	0,96	0,49	0,90
Random Forest	0,92	0,93	0,70	0,98	0,41	0,91

Tanto durante la fase de entrenamiento y como de testing se obtienen métricas muy similares lo que nos permite garantizar la existencia de consistencia en los modelos durante ambas fases. Para este experimento, Random Forest muestra mejores métricas de desempeño. Las predicciones obtenidas con Random Forest se detallan a continuación:

Prediction	0	1	All
True			
0	10729	243	10972
1	815	570	1385
All	11544	813	12357

- **Experimento 1 vs Experimento 2**

Las métricas de desempeño de Random Forest para ambos experimentos son muy similares. Por tanto se recomienda el Experimento 2 con el modelo Random Forest pues es más simple y se ejecuta más rápido. Asimismo al ser un modelo con menos variables, se facilita la tarea de manejo y almacenamiento de la información.

5. Conclusiones

- La tasa de aceptación de los depositos a plazo es baja con respecto a la tasa de rechazo (11% vs 89%).
- Proporcionalmente un 15% de las personas que se contactan por celular aceptan el deposito.Un 5% de las personas que son contactadas por teléfono aceptan el deposito.

- Hay una diferencia significativa en la tasa de aceptación del deposito dependiendo del nivel educativo.
- En promedio la aceptación de depositos a plazo diaria es de un 11% .
- El mes donde se realizan más llamadas es mayo, sin embargo este mes es el que muestra una de las tasas de aceptación más bajas con un 6%. Proporcionalmente los meses de marzo, septiembre, octubre y diciembre son los que proporcionalmente muestran mejores tasas de aceptación.
- Las personas casadas son a las que se contactan con mayor frecuencia, sin embargo los clientes solteros y estatus desconocido son los que proporcionalmente mostraron mayor tasa de aceptación de los depositos a plazo.
- Proporcionalmente los clientes mayores a 70 y entre 60-69 son los que muestran mejores tasas de aceptación, 47% y 43% respectivamente.
- A pesar de que los clientes bajo un perfil de estudiante y retirado no están entre los más contactados, proporcionalmente sus tasas de aceptación son de las más altas con un 31% y 25% respectivamente.
- Las personas que aceptaron el deposito a plazo fueron contactados menos veces en comparación con los que lo rechazaron.
- Aquellos clientes que aceptaron el deposito a plazo registraron duraciones de llamadas mayores con respecto a los que rechazaron el deposito.
- El algoritmo clasificatorio Random Forest presentó mejores métricas de desempeño en ambos experimentos realizados.

6. Recomendaciones

- Las métricas de desempeño de Random Forest para ambos experimentos son muy similares. Por tanto se recomienda el Experimento 2 con el modelo Random Forest pues es más simple y se ejecuta más rápido.
- Capturar otros atributos que puedan afectar la variable dependiente y el telemercadeo tales como características específicas del producto.
- Sería importante enriquecer el modelo mediante la incorporación de la variable año y hacer análisis más preciso del comportamiento de los depositos a plazos a través del tiempo. También sería interesante dividir la muestra de acuerdo con dos subperíodos de tiempo dentro del rango de 2008-2012, lo que nos permitiría analizar el impacto de la recuperación después de la recesión.
- Realizar las estrategias de telemercadeo enfocados en el tipo de cliente, un análisis de clustering sería muy valioso y complementario a lo realizado en este estudio.
- La calidad de la data es directamente proporcional a la calidad de los resultados a obtener, por estar razón es importante evitar niveles dentro de las variables como “unknown”. Nivel educativo y estado civil muestran estos niveles desconocidos y poco específicos.