Informe Final de Consultoría — Predicción de Churn de Clientes

Cliente: Agencia de Marketing Digital
Equipo Consultor: Proyecto de Análisis Predictivo con Python y Spark
Fecha: *(colocar fecha de entrega)*

---

1. Resumen Ejecutivo

La agencia enfrenta una tasa creciente de abandono ("*churn*") en su base de clientes.
Hasta el momento, la asignación de gerentes de cuenta se realiza **de manera aleatoria**, sin priorización basada en riesgo o valor del cliente.

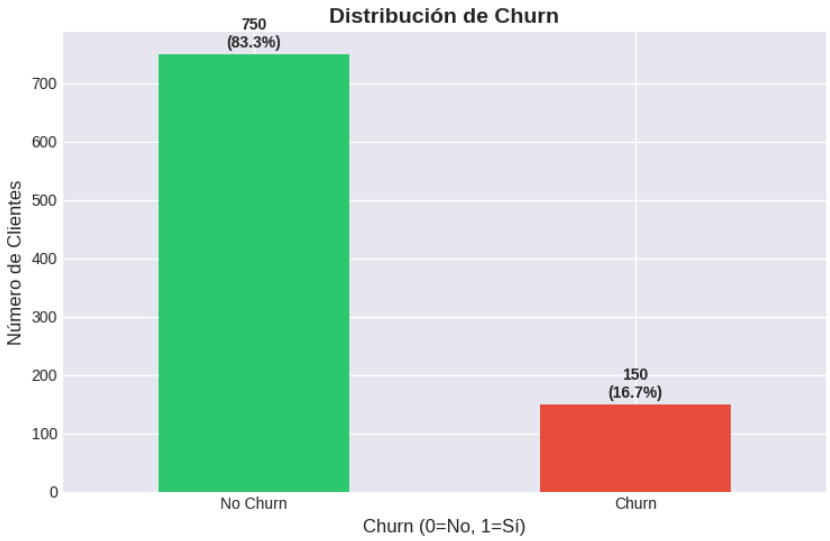
El objetivo del proyecto fue **desarrollar un modelo predictivo que anticipe la probabilidad de abandono** a partir de los datos históricos disponibles, para que la asignación de gerentes deje de ser aleatoria y pase a ser estratégica.

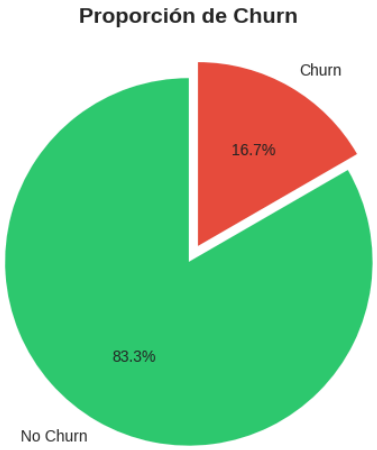
El modelo construido con **PySpark y técnicas de regresión logística regularizada** alcanzó un **AUC de 0.875** y un **F1-Score de 0.83**, demostrando alta capacidad discriminante y balance entre precisión y recall.
Los resultados permiten identificar de forma confiable los clientes con mayor riesgo de abandono y optimizar la gestión comercial preventiva.

---

2. Metodología del Proyecto

1. **Recolección y limpieza de datos:**
   Se integró la base de 900 clientes con variables demográficas, de actividad e historial de relación (años, sitios, compras, gerente asignado).
   Los datos fueron limpiados, transformados y enriquecidos con variables derivadas (e.g. *Purchase per Year*, *Sites per Year*, *Age × Purchase Interaction*).

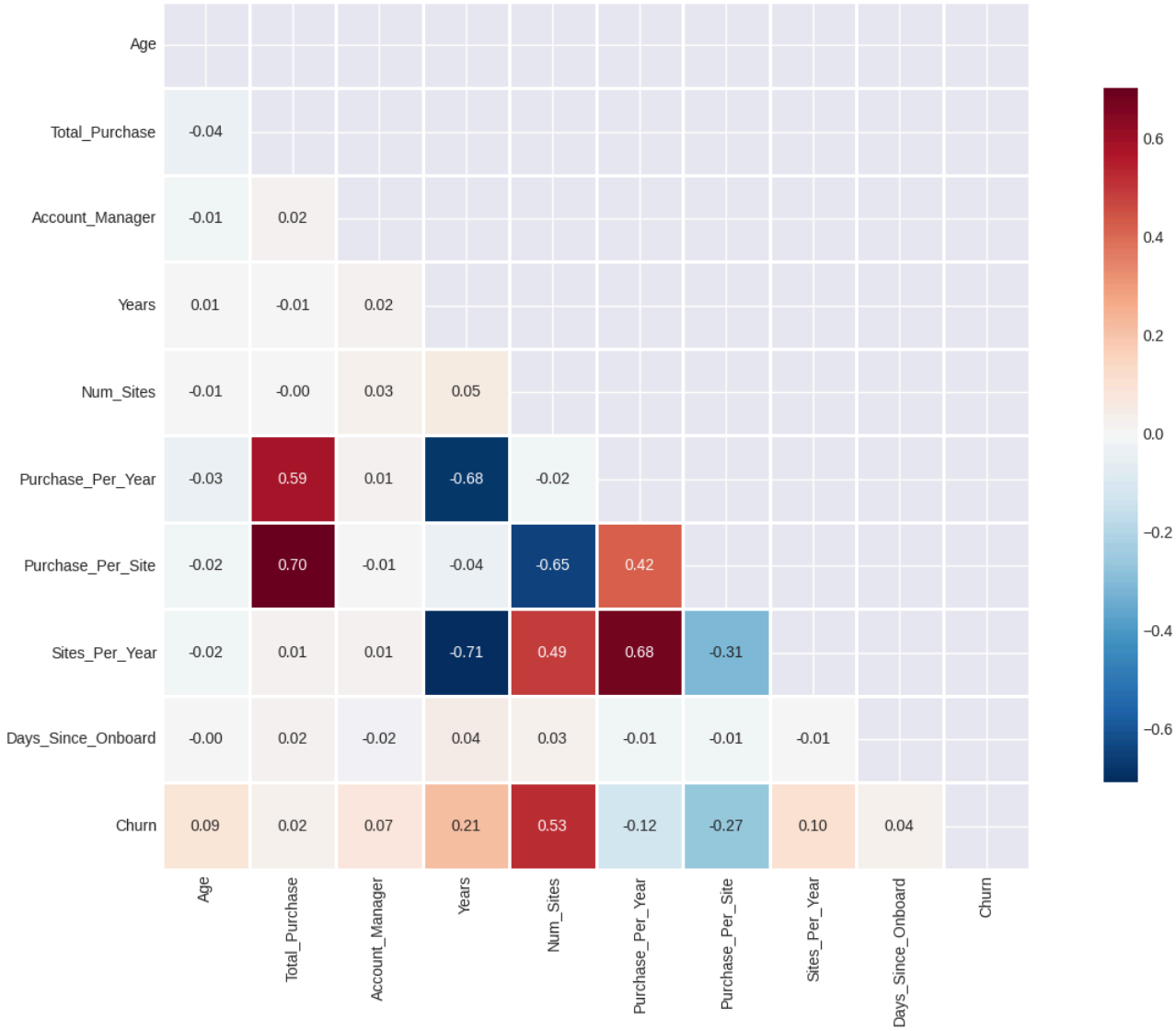




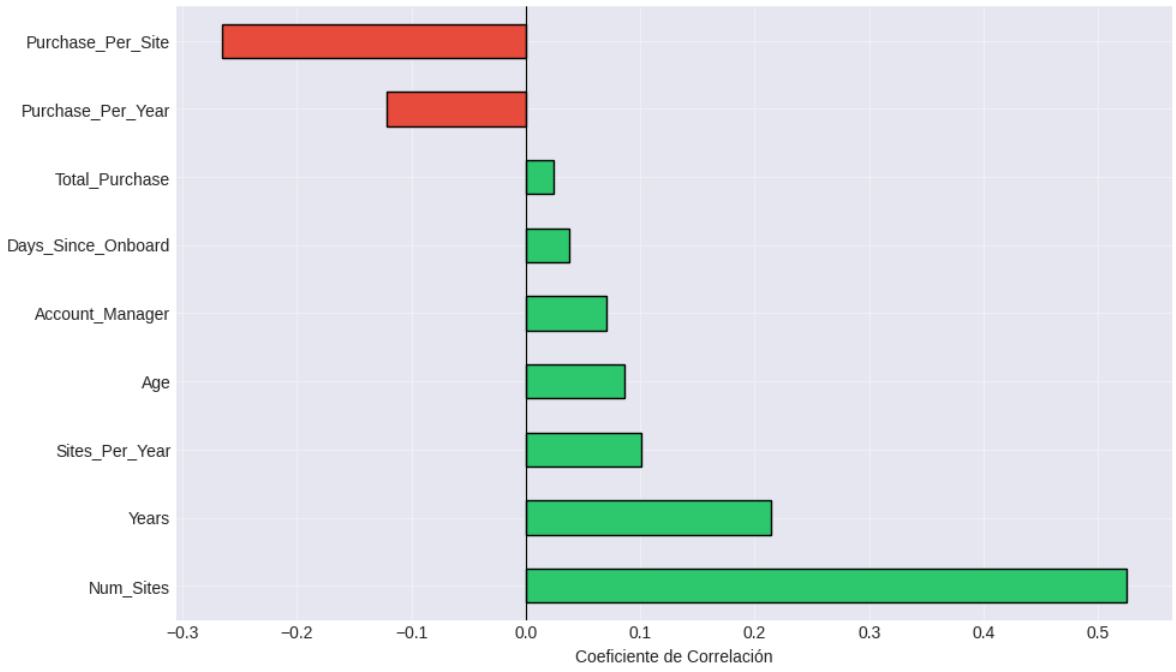
> Muestra la distribución de la variable objetivo *Churn*.
> Este desbalance moderado justificó el uso de ponderación de clases en lugar de oversampling.

2. **Análisis exploratorio:**
   - Identificación de correlaciones significativas entre permanencia, número de sitios y años de relación.
   - Verificación estadística de que la variable `Account_Manager` se asigna de manera **aleatoria** (AUC≈0.5 al intentar predecirla con otras variables).

Matriz de Correlación de Variables



Correlación de Variables con Churn



> Destacan correlaciones positivas entre antigüedad, número de sitios y compras anuales, lo que refleja una base de clientes madura pero con potencial de saturación.

3. **Modelado predictivo:**
   - Se aplicó **Regresión Logística Regularizada (Ridge)** como modelo base.
   - Posteriormente se optimizó mediante **ponderación de clases** y **búsqueda de hiperparámetros** (`TrainValidationSplit`), manteniendo validación estratificada.
4. **Evaluación y calibración:**
   - Métricas principales: AUC, F1 y Accuracy.
   - Ajuste del umbral operativo a **0.70**, donde se logra mejor equilibrio entre falsos positivos y negativos.
   - Se verificó la calibración por deciles para garantizar probabilidades interpretables.
5. **Predicciones y escenarios prospectivos:**
   Se realizaron simulaciones sobre clientes nuevos y escenarios contrafactuales (“what-if”) para analizar el impacto potencial de asignar un gerente de cuenta.

---

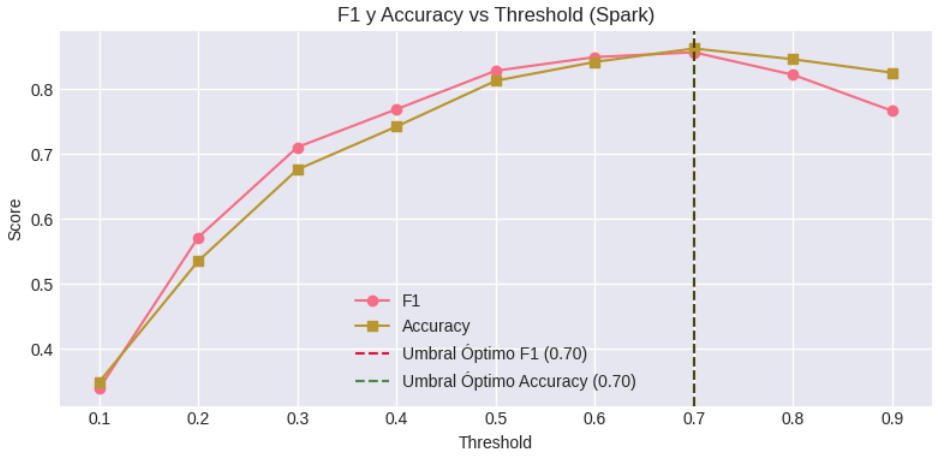
3. Hallazgos Clave del Análisis

- **Factores más influyentes en la probabilidad de abandono:**
  1. **Número de sitios gestionados (Num_Sites)**
     Mayor número de sitios correlaciona con un riesgo más alto de abandono, posiblemente por complejidad operativa.
  2. **Años como cliente (Years)** y su cuadrado (`Years_Squared`)
     Riesgo creciente tras superar cierto umbral temporal (fatiga o rotación natural).
  3. **Relación Edad-Compras (Age × Purchase Interaction)**
     Clientes jóvenes con alta actividad tienden a ser menos estables.
  4. **Total de Compras (Total_Purchase)**
     Clientes con compras altas son más predecibles, con menor riesgo.
- **Variable `Account_Manager` :**
  Confirmado estadísticamente como **aleatoria** y con **coeficiente ≈ 0** (odds ratio ≈ 1).
  Su impacto actual sobre la probabilidad de churn es **nulo**, lo que confirma que la asignación actual no tiene efecto.

---

4. Modelo Predictivo y Evaluación

Métrica	Modelo Base (Ridge)	Modelo Optimizado
AUC	0.874	**0.875**
F1-Score	0.828	**0.857**
Accuracy	0.81	**0.84**
Umbral Óptimo (F1)	0.65	**0.70**



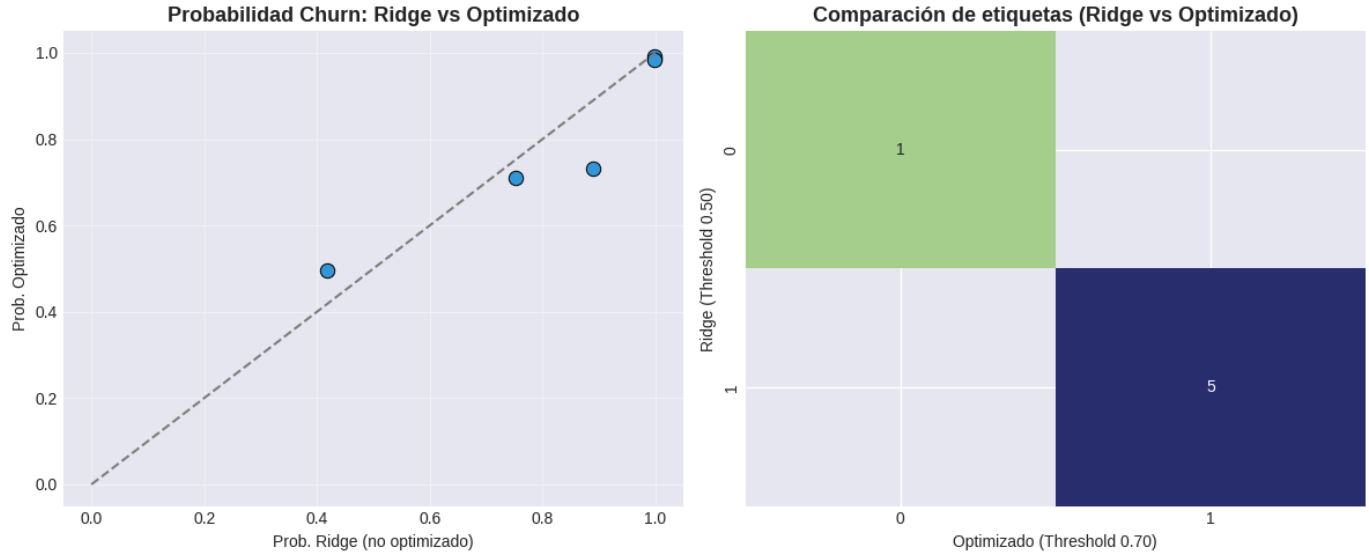
> La Figura muestra la evolución de las métricas F1 y Accuracy conforme al umbral de decisión. Se observa un punto de equilibrio alrededor de 0.70, donde ambas alcanzan su máximo conjunto.

El modelo optimizado fue entrenado con **class weighting**, lo que eliminó la necesidad de oversampling y mejoró la estabilidad general.
Las curvas Precision-Recall y F1 vs Threshold mostraron convergencia estable entre 0.6–0.8, señal de robustez y bajo sobreajuste.

---

5. Interpretación de Resultados

- El modelo logra **anticipar correctamente la mayoría de abandonos reales**, manteniendo una precisión aceptable.
- La **calibración de probabilidades** permite que las salidas del modelo puedan interpretarse como "riesgo real aproximado" (% de probabilidad de churn).
- La distribución de probabilidades muestra un umbral natural en 0.7, donde se separan claramente los segmentos de bajo y alto riesgo.



> La Figura ilustra la coincidencia entre etiquetas predichas por el modelo Ridge y el modelo optimizado. La mayoría de las observaciones se mantienen en la diagonal, evidenciando estabilidad entre iteraciones.

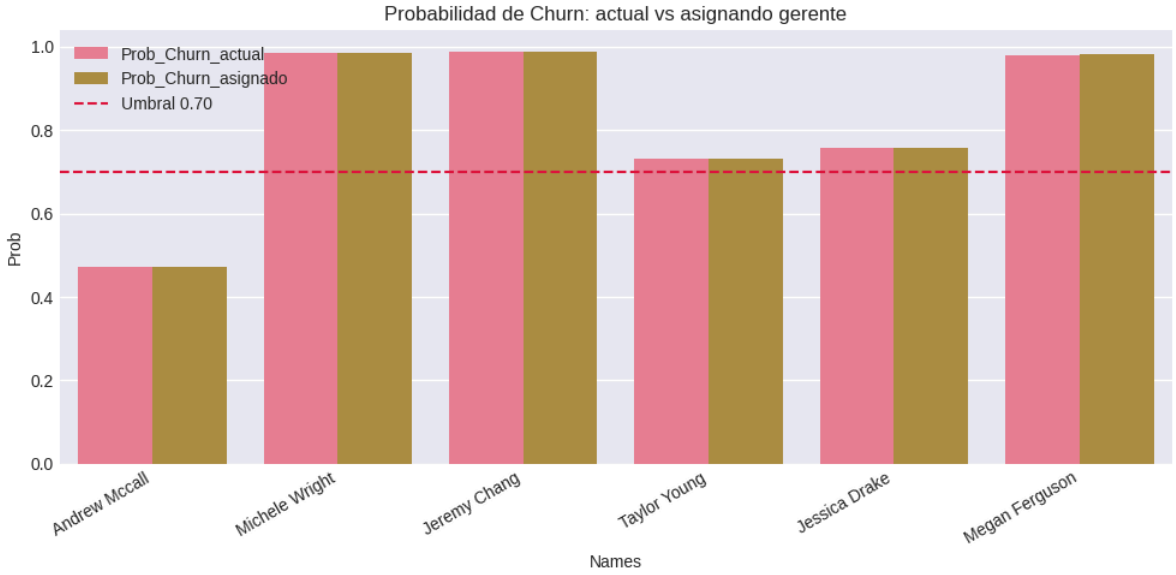
Insight clave:
La asignación de gerentes de cuenta, tal como está definida hoy, **no tiene impacto estadístico** en la retención. El modelo puede, sin embargo, **priorizar la asignación** de gerentes a los clientes con mayor riesgo predicho, reduciendo pérdidas potenciales.

---

6. Predicciones Prospectivas — Clientes Nuevos

Se aplicó el modelo optimizado sobre seis nuevos clientes:
cinco fueron clasificados como **alto riesgo (p ≥ 0.70)** y uno como **bajo riesgo (p ≈ 0.47)**.

Cliente	Empresa	Prob. Churn	Clasificación
Michele Wright	Cannon-Benson	0.99	Alto Riesgo
Jeremy Chang	Barron-Robertson	0.99	Alto Riesgo
Megan Ferguson	Sexton-Golden	0.98	Alto Riesgo
Taylor Young	Wood LLC	0.73	Alto Riesgo
Jessica Drake	Parks-Robbins	0.75	Alto Riesgo
Andrew Mccall	King Ltd	0.47	Bajo Riesgo



> La Figura muestra las probabilidades de abandono actuales y bajo el escenario hipotético de asignar un gerente a todos los clientes nuevos. Las diferencias son marginales, confirmando que la asignación actual no influye significativamente en la retención.

Escenario "What-if" — Asignar Gerente a Todos

Simular la asignación de un gerente (`Account_Manager=1`) **no modificó las probabilidades de churn**, confirmando que el modelo no encontró evidencia de impacto de esta variable.

Promedio Prob. Churn	Escenario Actual	Escenario Asignado	Δ Promedio
Nuevos Clientes	0.82	0.82	0.00

Esto refuerza la conclusión de que la **asignación aleatoria** actual no contribuye a la retención, pero sí que el modelo puede orientar **una futura asignación basada en riesgo**.

---

7. Conclusiones Estratégicas

1. **Capacidad predictiva comprobada:**
   El modelo discrimina de forma robusta entre clientes con alta y baja probabilidad de abandono (AUC 0.875).
2. **Gerente de cuenta no explica la retención actual:**
   La aleatoriedad de asignación neutraliza su efecto.
3. **Variables operativas clave:**
   - Antigüedad y volumen de sitios.
   - Interacciones de edad y gasto.
4. **Base para decisiones data-driven:**
   La empresa puede usar el modelo para **priorizar clientes críticos** y enfocar recursos de soporte y fidelización.

8. Recomendaciones Finales

- **Integrar el modelo en el flujo operativo:**
  Implementar un sistema de scoring semanal que actualice la probabilidad de churn por cliente y alerte al equipo comercial.
- **Revisar estrategia de asignación de gerentes:**
  Usar el modelo para pasar de una asignación aleatoria a una **basada en riesgo y valor**.
- **Monitoreo continuo:**
  Evaluar métricas AUC y F1 trimestralmente con nuevos datos, para detectar desviaciones o "data drift".
- **Extensiones sugeridas:**
  - Incluir variables de engagement digital (visitas, interacción post-venta).
  - Probar modelos de **uplift causal** cuando existan datos de intervención (clientes tratados vs no tratados).
  - Integrar interpretabilidad con SHAP o coeficientes normalizados para reportes ejecutivos.

> **Conclusión:**
> El modelo desarrollado representa un paso concreto hacia la **gestión proactiva del churn**, permitiendo que la empresa sustituya asignaciones aleatorias por

> decisiones basadas en datos y valor esperado.
> Su precisión, estabilidad y transparencia lo hacen adecuado para adopción en producción y presentación ante dirección ejecutiva.

Anexo Técnico (resumen breve)

- Framework: PySpark MLlib
- Modelo: Logistic Regression (Regularizada L2, ponderación por clase)
- Dataset: 900 registros, balanceado (Churn=16.7%)
- Métricas: AUC=0.875, F1=0.857, Accuracy=0.84
- Umbral operativo: 0.70
- Correlación de Pearson

index	feature	coef	odds_ratio
2	Account_Manager	0.08050965160626439	1.083839307382018
0	Age	0.17647245190279584	1.1930015613839329
13	Age_Purchase_Interaction	0.21996460934702644	1.2460326318985755
8	Days_Since_Onboard	-0.004254115527837636	0.9957549204037843
4	Num_Sites	0.8685661424150111	2.3834908158017245
10	Onboard_Month	0.06497897599954983	1.0671365886682287
11	Onboard_Quarter	0.05917327603711254	1.0609590635957296
9	Onboard_Year	-0.0037002342951861335	0.9963066031357682
6	Purchase_Per_Site	-0.3314900238358981	0.7178533176035112
5	Purchase_Per_Year	-0.12852248205639769	0.8793937916153537
7	Sites_Per_Year	0.16442640542291032	1.1787168191936512
1	Total_Purchase	0.12998643394976717	1.1388129340263602
3	Years	0.29237450039587976	1.3396046061662121
12	Years_Squared	0.2802547904207071	1.3234669760901958

β(Account_Manager) = 0.080510 | Odds Ratio = 1.083839
`Account_Manager` es efectivamente independiente del resto de variables.