BUSINESS CASE SOLUTION

Laureate International Universities

Gabriel Sainz Vázquez

Fecha: 13 de noviembre del 2022

Contents

Business Case

Eres un(a) consultor(a) responsable de ofrecer respuestas a miembros de una organización. La compañía busca entender cuáles de sus empleados son más propensos a abandonar la compañía (incurrir en attrition).

El objetivo es el siguiente:

- Generar un modelo que pueda ser utilizado por la compañía.
- Entender cuáles son las variables que tienen mayor impacto en la tasa de abandono de empleados.
- Proponer una estrategia para disminuir la tasa de abandono. Realizar sugerencias para medir la eficacia de la estrategia.

Los datos disponibles con los que podrás trabajar se encuentran en el archivo **Employee Attrition.csv**.

Solution

La solución del business case fue desarrollado de acuerdo con los siguientes puntos:

- **Descriptive Analysis**
  - Target Variable Analysis
  - Categorical Variables Analysis
  - Continuous Variables Analysis
- **Pipeline**
  - Missing Treatment
  - Categorical Features Treatment
  - Train Test Split
  - Variable Reduction
- **Modeling**
  - Logistic Regression Model
  - Balanced Method with Logistic Regression
  - XGBoost (Machine Learning Model)
  - Metrics
- **Best Model**
- **Strategies**

El script fue desarrollado en Python. En el siguiente link, se puede encontrar el código.

https://colab.research.google.com/drive/1dtmGDIeVChXG6HwosBlVeyu7ZJBfs_PN?usp=sharing

Descriptive Analysis

Primero, se desarrolló un análisis descriptivo:

Target Variable Analysis

De acuerdo con la base de datos, existe un 16.12% de registros considerados attrition. A continuación, podemos ver una gráfica de barras que representa que las clases están desbalanceadas.

Las clases desbalanceadas jugarán un papel importante a la hora del modelado de la información, ya que, algunos modelos pueden estar sesgados por la poca cantidad de buenos o malos en la muestra.

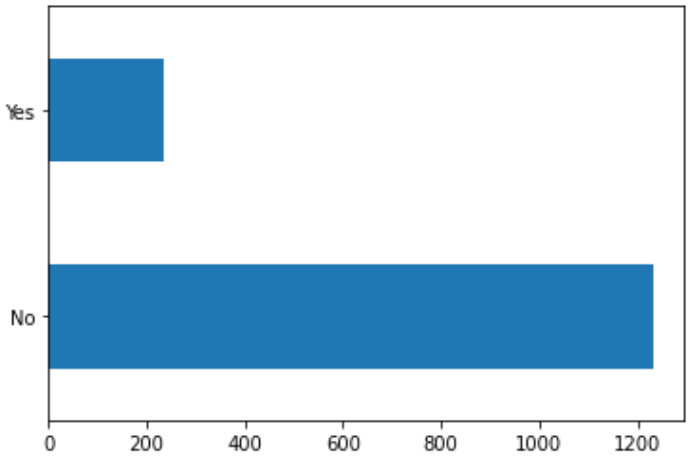


Figure 1. Attrition Employees

Posteriormente, un análisis descriptivo de acuerdo con las variables independientes fue hecho. Es importante incluir cómo se ven las diferentes variables en relación a la target para poder discriminar y ver gráficamente si alguna de las variables tiene más peso a la hora de predecir Attrition de empleados.

Categorical Variables Analysis

Para analizar conjuntamente las variables categóricas incluyendo la variable objetivo, se hizo un porcentaje de cuantos Attrition se tienen en cada una de las categorías.

En la Figura que se muestra a continuación, podemos ver las variables que tuvieron un porcentaje alto en ciertas categorías, lo cual, discriminan a los empleados que abandonaron el empleado y se relaciona directamente con las variables que tienen mayor impacto en la tasa de abandono.

Estas variables se enumeran a continuación:

- Business Travel
- Education Field
- Environment Satisfaction
- Job Involvement

- Martial Status
- Number Companies Worked
- Worked Over Time
- Training Time Last Year (Weeks)
- Work Life Balance

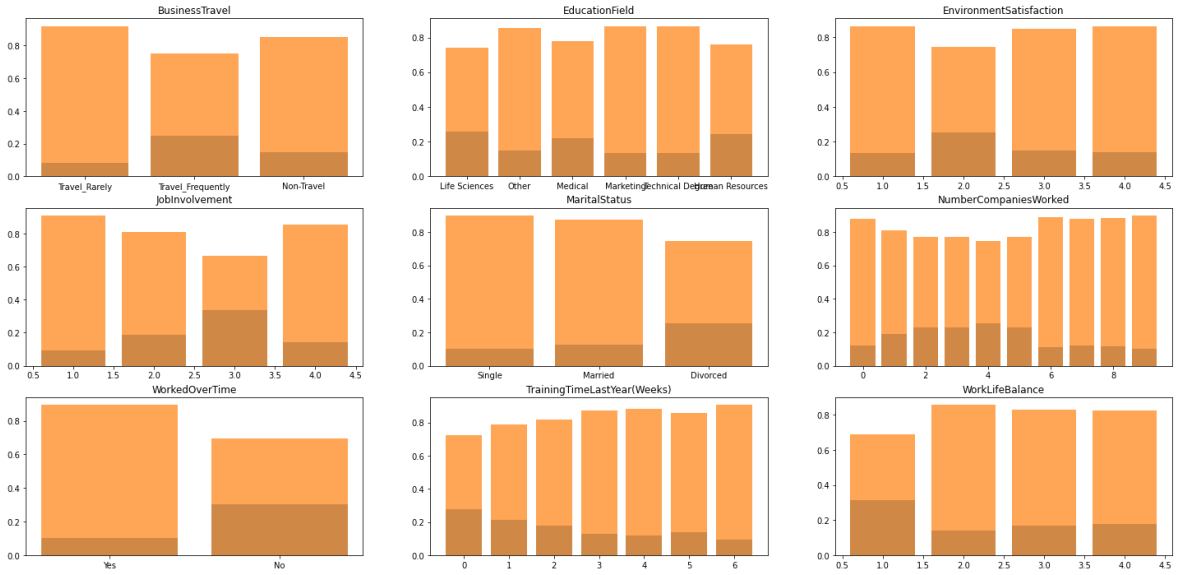


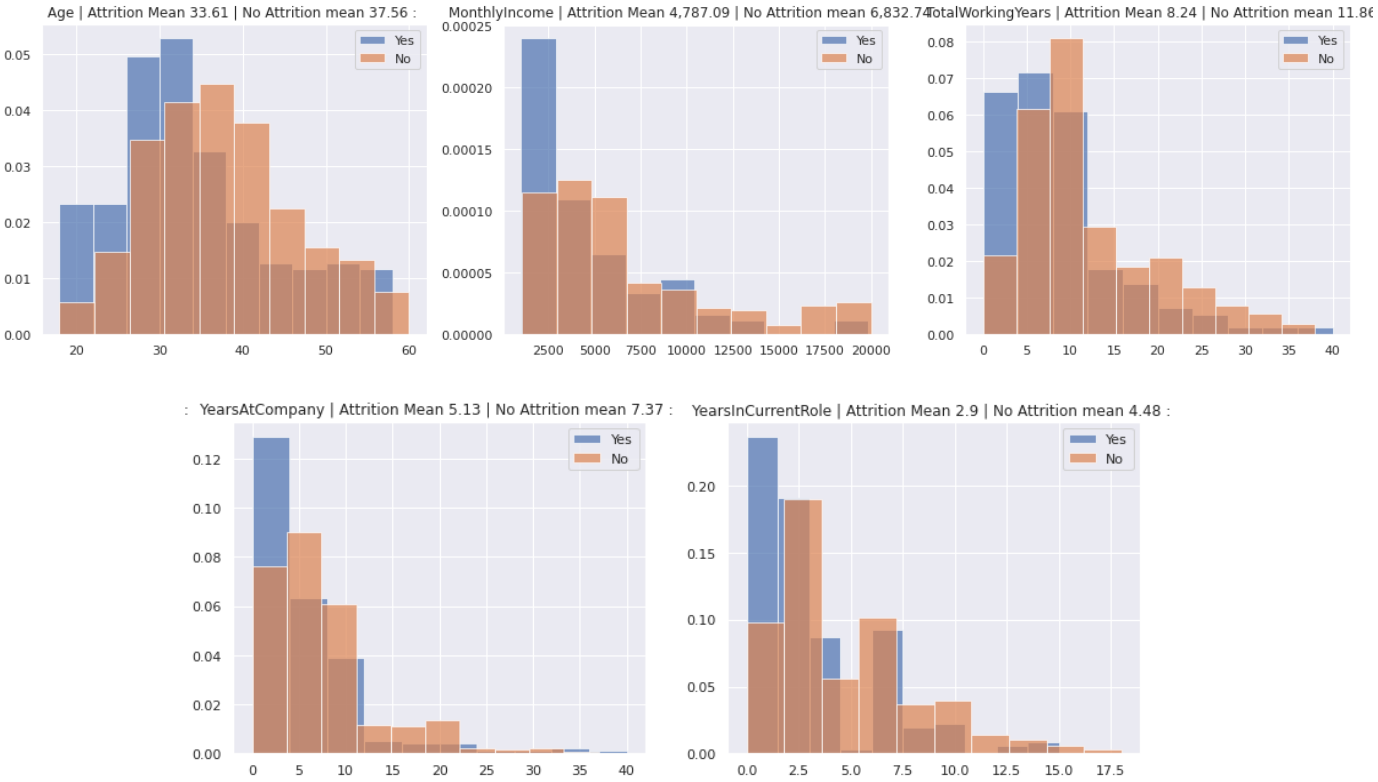
Figure 2. Significant Categorical Variables

Continuous Variables Analysis

De igual manera, para las variables continuas, se realizaron histogramas para ver si existe una diferencia significativa a la hora de discriminar por attrition.

Las siguientes variables tuvieron una media distinta de acuerdo con si fueron attrition o no:

- Age
- Monthly Income
- Total Working Years
- Years At Company
- Years In Current Role



Pipeline

El procesamiento de los datos se realizó como se muestra a continuación:

Missing Treatment

No se tienen missing en la información, por lo tanto, no se aplicó ningún tratamiento de missing treatment.

Categorical Features Treatment

Dado que la mayor parte de las variables son categóricas, se realizó un tratamiento para poder modelar numéricamente.

Se realizó “One Hot Encoding”. Esta estrategia, crear una columna para cada valor distinto que exista en la característica que estamos codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0.

Train Test Split

Uno de los objectivos del aprendizaje supervisado es la creación de modelos que puedan tener un buen performance en nueva data. Por lo tanto, el train/test Split nos ayuda a validar nuestro modelo en una muestra “test”.

En este ejercicio, se realizó un train test Split de 80/20.

Variable Reduction

Se utilizó el criterio del information value como método de reducción de variables.

El information value es una técnica que ayuda a determinar qué columna en el data set tiene un buen poder predictivo o influencia en el valor especificado como variable dependiente.

El information value se calcula de la siguiente manera:

$$IV = \sum_{i=0}^{n-1} WOE * (\%\, of\ non - evengs - \%\, of\ events)$$

$$WOE = \ln\left(\frac{\%of\ non\ events}{\%of\ events}\right)$$

Donde

- n es usualmente 10 (deciles).

Por lo tanto, entre más grande el Information Value, significa que existe más poder predictivo.

Normalmente, se considera un treshold del 0.02 para considerar que una variable tiene un buen poder predictivo.

De acuerdo con la metodología anterior, en el siguiente archivo podemos ver las variables con su respectivo information value.

Information_value.csv

Por lo tanto, con un treshold del 0.02, las variables óptimas se reducen a 31 variables.

Modeling

Para el modelado de la información, se utilizaron tres estrategias distintas:

- Regresión Logística
- Regresión Logística con método de balanceo
- XGBoost

Regresión Logística

Se realizó un entrenamiento del modelo con una regresión logística como benchark para comparar otros modelos.

Los resultados se muestran a continuación:

	precision	recall	f1-score	support
0	0.85	0.99	0.91	249
1	0.25	0.02	0.04	45
accuracy			0.84	294
macro avg	0.55	0.51	0.48	294
weighted avg	0.76	0.84	0.78	294

```
array([[246,   3],
       [ 44,   1]])
```

A pesar de que la precisión es muy buena, la métrica f1 está muy baja. Esto se debe al desbalanceo de los datos.

Está prediciendo casi a todos los empleados como No attrition. Por lo tanto, para esta metodología es necesario un balanceo de la muestra.

Regresión Logística con Método de Balanceo

Se entrenó un modelo con una metodología que balancea los datos.

Las métricas se muestran a continuación:

```
The f1 score for the testing data: 0.47058823529411775
```

	precision	recall	f1-score	support
0	0.94	0.76	0.84	249
1	0.35	0.71	0.47	45
accuracy			0.76	294
macro avg	0.64	0.74	0.66	294
weighted avg	0.85	0.76	0.78	294

```
[[190  59]
 [ 13  32]]
```

Como podemos ver, las métricas mejoraron. A pesar de que el accuracy es de 0.76, la métrica f1 mejoró significativamente.

XGBoost

Este método se basa en árboles de decisión y supone una mejora sobre otros métodos, como el bosque aleatorio y refuerzo de gradientes.

Para este modelo, no es necesario realizar un balanceo de muestras. Existen artículos científicos que comentan que no existe una diferencia significativa que ayude el balanceo de datos, ya que, se basa en árboles de decisión.

Se realizó un hyperparamether tuning para encontrar al mejor modelo.

La configuración del grid de hiperparametros fue la siguiente:

```
params = {'n_estimators': [10, 15, 20, 25, 30],
          'learning_rate': [0.01, 0.05, 0.1, 0.15, 0.25, 0.35],
          'max_depth': [3,4,6,8,10,15],
          'colsample_bytree': [0.6,0.7,0.8],
          'subsample': [0.7,0.8,1],
          'min_child_weight': [1,2,4]}
```

Cada modelo con sus respectivas métricas, los podemos ver en el siguiente documento:

metrics_models.csv

Posteriormente, el mejor modelo se presenta a continuación:

Best Model

```
BEST MODEL
{'colsample_bytree': 0.8, 'learning_rate': 0.35, 'max_depth': 3, 'min_child_weight': 4, 'n_estimators': 25, 'subsample': 0.8}
       The f1 score | Train: 0.7055 | Test: 0.54286
       Accuracy score | Train: 0.92262 | Test: 0.89116
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       249
           1       0.76      0.42      0.54        45

    accuracy                           0.89       294
   macro avg       0.83      0.70      0.74       294
weighted avg       0.88      0.89      0.88       294

[[243   6]
 [ 26  19]]
```

El mejor modelo es un XGBoost y tienen los siguientes hiperparámetros:

- 'colsample_bytree': 0.8
- 'learning_rate': 0.35
- 'max_depth': 3
- 'min_child_weight': 4
- 'n_estimators': 25
- 'subsample': 0.8

Podemos ver que el accuracy es de 0.89, con un f1 score de 0.54.

Por lo tanto, es el mejor modelo en comparación con las regresiones logísticas.

Como podemos ver, el mejor modelo cuenta con 25 estimadores, los cuales son los siguientes con la importancia en el modelo:

feature_importance.csv

Las 10 variables con mayor relevancia se muestran a continuación:

	Columns	realtive_importance	cumulative_importance
5	JobLevel	0.076466	0.076466
6	TotalWorkingYears	0.070340	0.146806
4	WorkedOverTime	0.064283	0.211089
11	JobInvolvement	0.056553	0.267642
8	Single	0.052030	0.319672
10	NumberCompaniesWorked	0.043376	0.363049
23	YearsSinceLastPromotion	0.040690	0.403738
1	YearsInCurrentRole	0.040373	0.444112
3	MonthlyIncome	0.040351	0.484463
15	Travel_Frequently	0.039890	0.524352

Podemos ver que sí se relaciona con el análisis descriptivo.

Strategies

De acuerdo con el análisis visto a continuación, los empleados que tienen attrition dependen mayormente de cómo están involucrados en el trabajo, estado civil, edad y cuanto tienen de ingresos, por lo tanto, podríamos suponer que son empleados jóvenes que no tienen mucho tiempo trabajando en este empleado.

Por lo tanto, se considerarían las siguientes propuestas:

- Crear incentivos para empleados nuevos y empleados con antigüedad, ya sea con bonos, horarios flexibles, promociones y un ambiente laboral que los haga sentir motivados día con día.

- Existe una gran cantidad de empleados jóvenes que buscan otras oportunidades, por lo tanto, crear un plan de carrera, considero que es fundamental en las primeras etapas de los trabajadores.
- Crear una sinergia que les de valor a los empleados.
- Brindar cursos para tener un aprendizaje continuo.

Por otro lado, para medir la eficiencia de la estrategia, se puede realizar un análisis mensual de la satisfacción del empleado. Crear cuestionarios simples y rápidos que nos ayude a determinar si el empleado está satisfecho.

De igual manera, seguir generando esta cantidad de datos para tener un análisis estadístico y ver el progreso de la estrategia.