

Unai Famoso Rodríguez

# **Presentación de técnica**

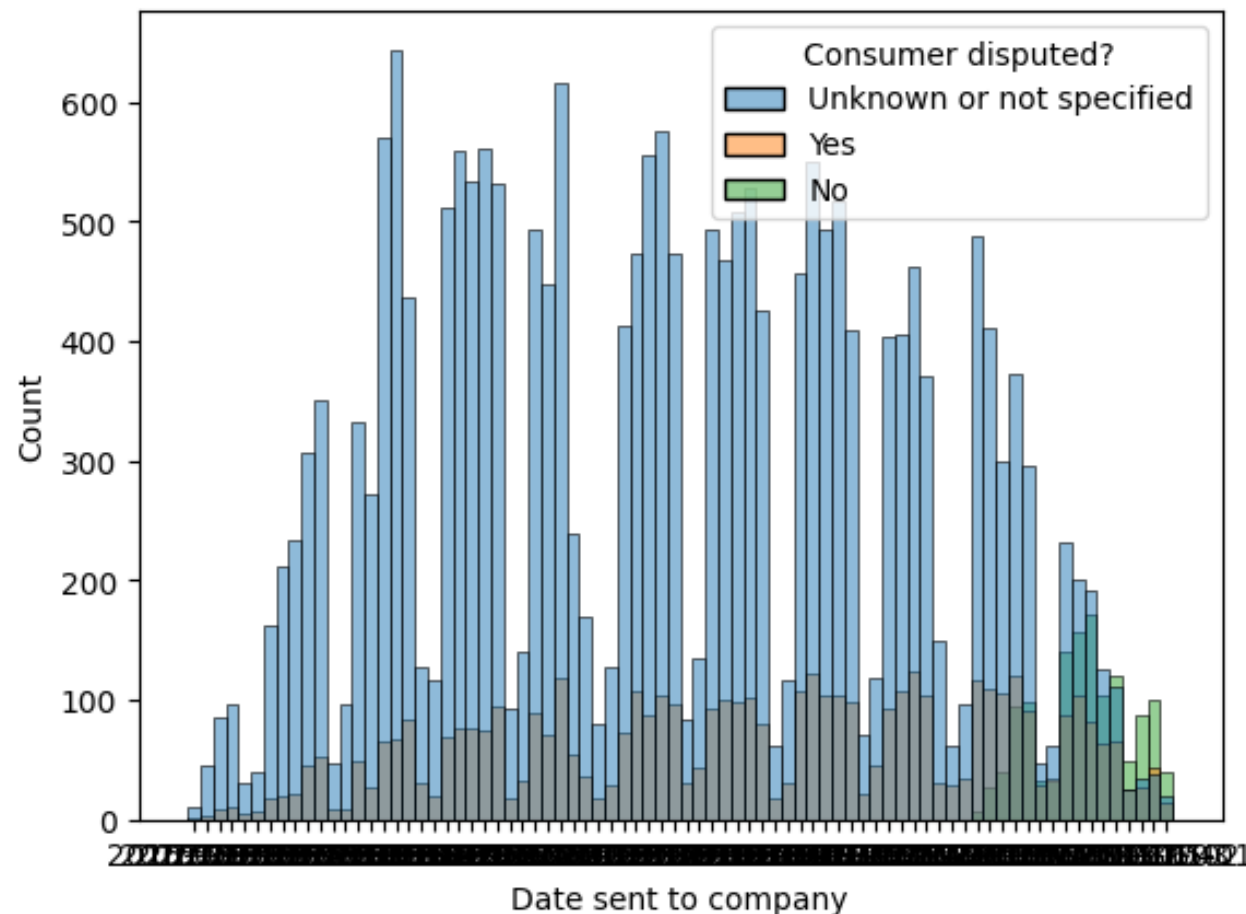
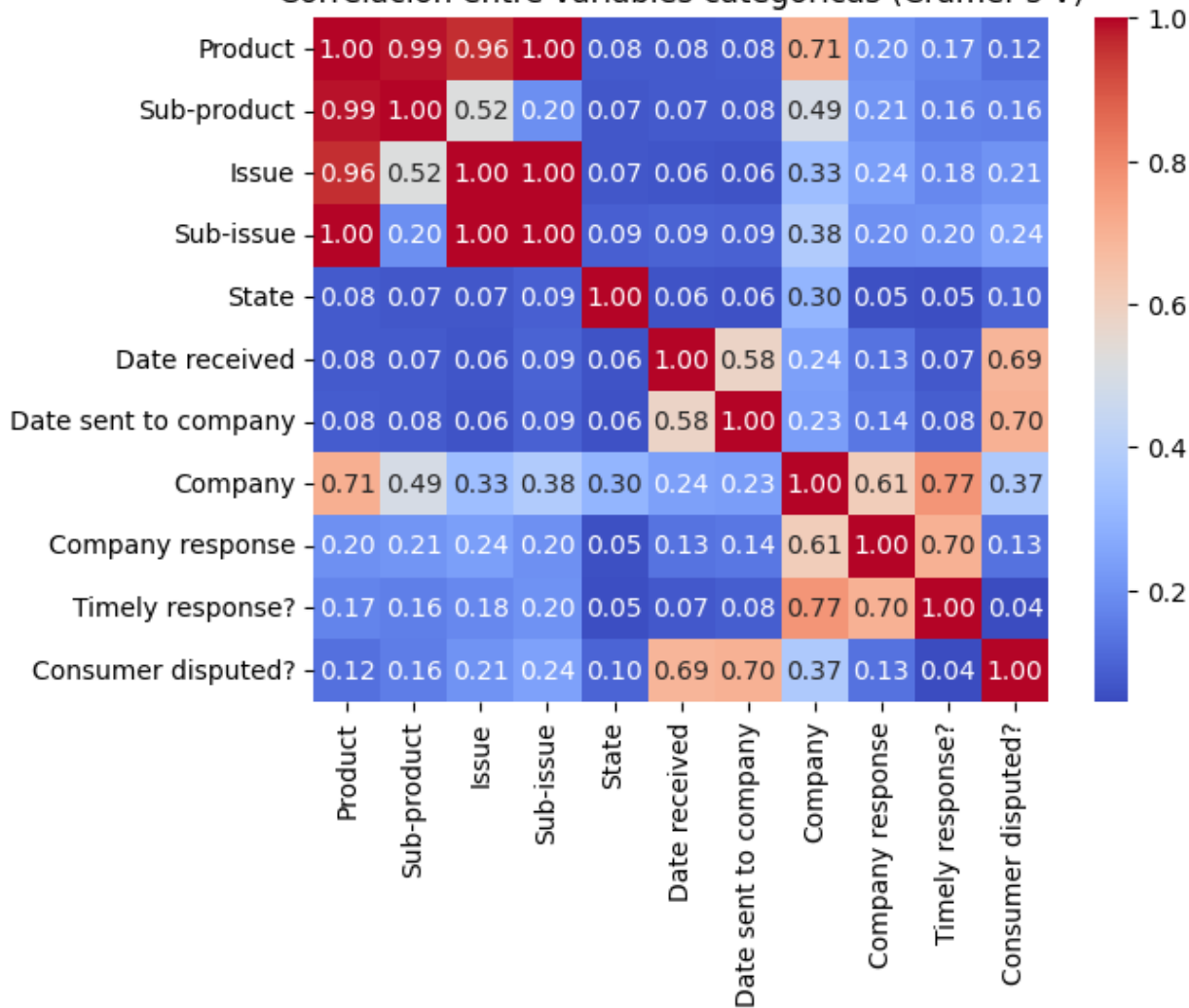


# Contextualización

- Se pretende estimar si:
  - la respuesta a la queja será a tiempo (modelos **timely**)
  - Si el cliente disputara o no (modelos **dispute**).
- Ambos son modelos de clasificación binaria.
- Tras esto se pretende utilizar ambos modelos e introducirlos en una app de *gradio* o *API* de *FAST API* como prototipo.

	Ej.	Rango	Vacios?
Complaint ID	1291006		No
Product	Hipoteca	11	No
Sub-product	Hipoteca asegurada	43	
Issue	Tácticas de comunicación	90	No
Sub-issue	Llamadas repetitivas o frecuentes	47	Si
State	TX	60	Si
ZIP code	76119.0	---	Si
Date received'	2015-03-19	Desde 2015-01-01 2015-03-19	No
Date sent to company	2015-03-19	Desde 2015-01-01 2015-03-19	No
Company response	Closed with explanation	6	No
Company	Premium Asset Services, LLC	1534	No
<b><u>Timely response?</u></b>	Si		No
<b><u>Consumer disputed?</u></b>	Si		Si

Correlación entre variables categóricas (Cramér's V)



# Limpieza y transformaciones en los datos

Para los casos en los que se tenía el *ZIP code* pero no el estado de procedencia se ha utilizado la API *zippotam*.

- Para los ZIP vacíos se relleno con el código ZIP 0
- Para los estados desconocidos se relleno con la etiqueta *Unknown*
- Para el resto de las variables se relleno con la etiqueta *Unknown or not specified*

Para *feature engineering* se sacaron 3 variables adicionales de las fechas:

- Días de retraso (que no se acabo usando porque empeora los modelos)
  - *Weekday* (En forma de texto el día que se recibió)
  - *Holiday* (si es día de fiesta o no el día que se recibió. No se acabo usando porque todos son laborables)
- 
- En todos los casos se eliminaron los casos con company response *untimely response* o *in progress*
  - En el modelo *timely* se realizo *label encoding* en dispute *one-hot encoding*
  - Se dividió el *dataset* 80:20 barajados

# Enfoque de la metodología timely

- Para el modelo se eliminaron
  - Los casos con Company response: *in progress*, y untimely response
  - Columna: Company
  - Columna: Company response
  - Columna: *Complaint ID*
  - Columna: *Consumer disputed?*
  - Columna: *ZIP code*
  - Columna : *Fechas (porque es un rango limitado)*
- *En feature engineering solo se saco el Estado si esta vacío*
- *Se codifico en label*
- *Se hizo resample con ADASYN para evitar desbalanceo*
- *Se dividió el dataset 80:20 barajados*



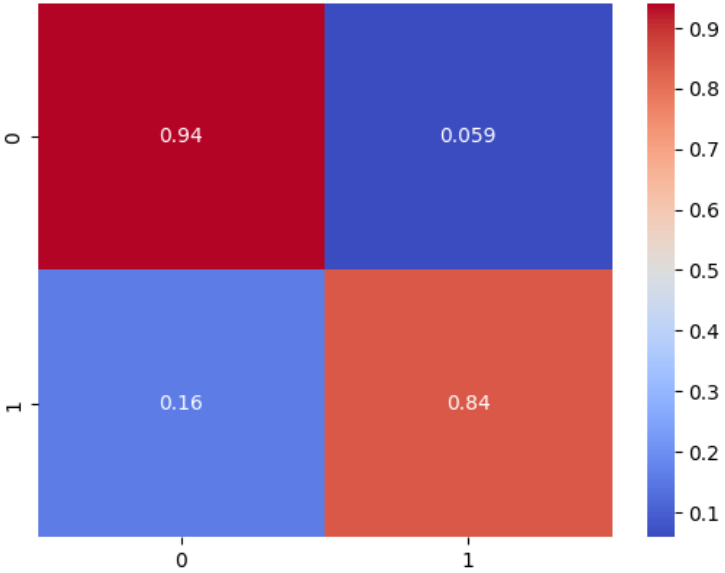
# Enfoque de la metodología timely

- El modelo que se ha elegido es un *tree classifier* optimizando con grid search el AUC-ROC score:
  - criterion
    - gini
    - entropy
  - max\_Depth= de 1 a 40
  - min\_samples\_leaf=de 15 a 30
- El modelo final
  - Criterion: entropy
  - max\_Depth 25
  - min\_samples\_leaf: 15

# Resultados y métricas de evaluación timely

- Buenos estadísticos y equilibrados
- Predice mejor los casos negativos
- No esta demasiado sobreajustado (0.89)
- Aunque modelos con mejores estadísticos
  - Mucho mas pesados
  - Sobreajustado
  - Menos interpretable

Feature Importance	
Sub-product	0.24
Product	0.24
State	0.22
Issue	0.16
Sub-issue	0.14

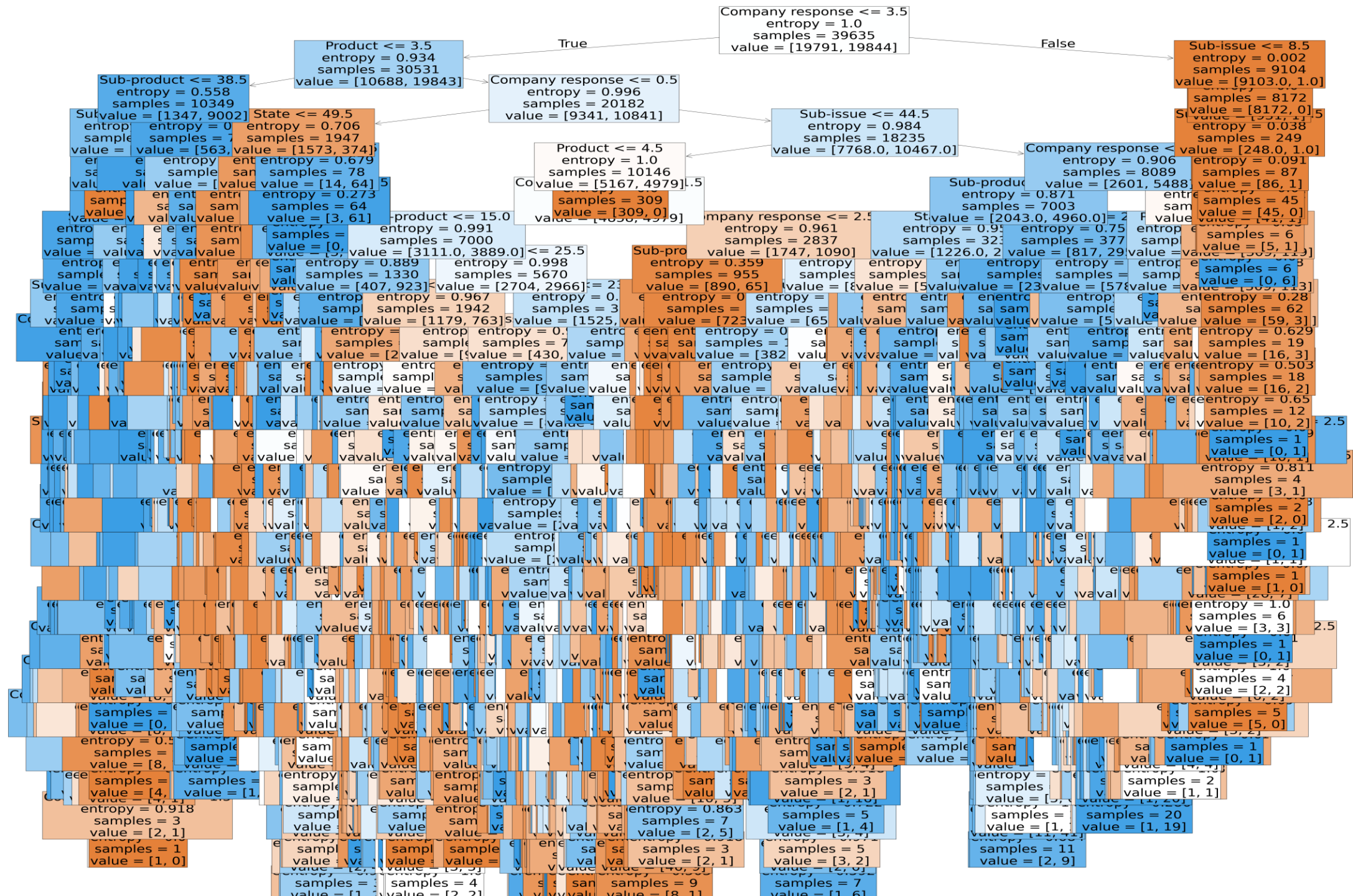


Clase	Precisión	Recall	F1-Score	Soporte
No	0.86	0.94	0.90	4990
Yes	0.93	0.84	0.89	4923
Accuracy Total			0.89	9913
ROC AUC			0.89	
Promedio Macro	0.90	0.89	0.89	9913
Promedio Ponderado	0.90	0.89	0.89	9913

# Discusión sobre limitaciones y mejoras *timely*

- No está en pipeline con preprocesamiento → dificulta implementación
- No se evaluaron todos los modelos posibles, es posible que se encuentren modelos mas adecuado
- No tiene una *accuracy* tan alta (0.89) para fiarse de ella
- Difícil de interpretar por su alta profundidad
- Es necesario optimizar el umbral para adecuarse a las necesidades de la empresa
- Seria mas útil se hubiese hecho un modelo de regresión (mas preciso)







# Enfoque en la metodología dispute

- El único modelo tipo de modelos que consiguió aprender algo fueron las redes neuronales
- Para el modelo se eliminaron
  - Los casos con Company response: *in progress*, y untimely response
  - Columna: *Complaint ID*
  - Columna: *ZIP code*
  - Columna : *Fechas (porque es un rango limitado)*
- Se hizo *resampling* un oversampler ADASYN para combatir el desbalanceo
- Se dividió el dataset 80:20 barajados

# Enfoque en la metodología dispute

- Arquitectura del modelo

1. *Input Layer*
2. Capa densa (*ReLU*, regularización l2 0.001)
3. Batch Normalización
4. *Dropout* (30%)
5. Capa densa (Sigmoidea)

**Capas ocultas**

256  
128  
64  
32

- Entrenamiento y optimización (maximizar *accuracy* y *AUC-ROC*)

- Batch 64
- Val 0.20
- 500 epoch

- Early Stopping (paciencia: 20)

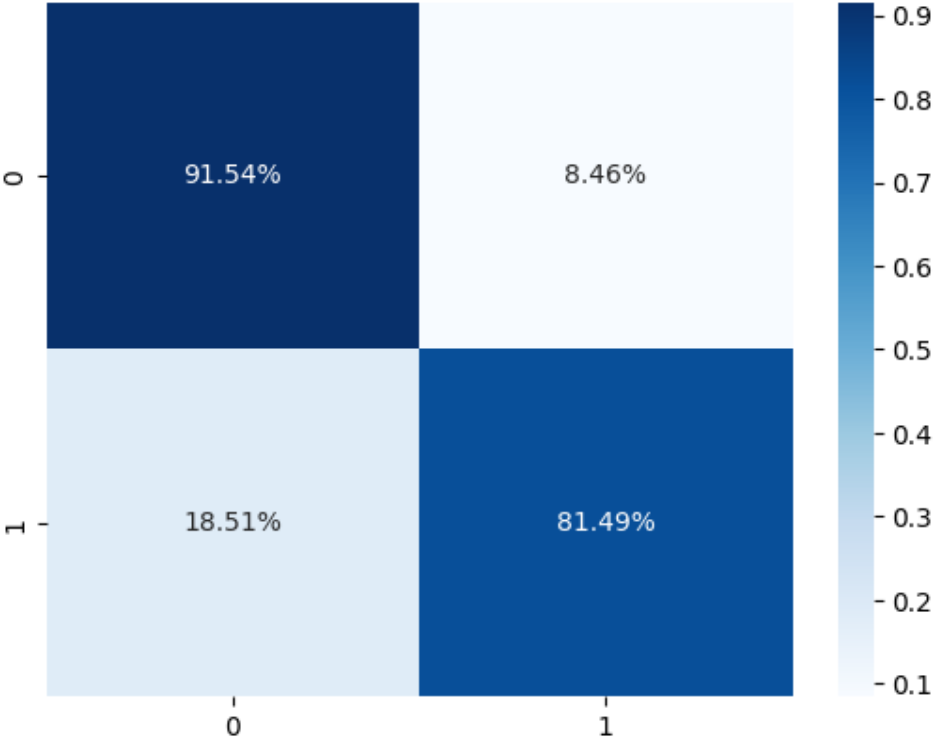
- ReduceLROnPlateau

- monitor = *val\_loss*
- factor = 0.5
- paciencia = 10
- min\_lr = 1e-6



# Resultados y métricas de evaluación dispute

- Tiene unos buenos estadísticos y bastante equilibrados

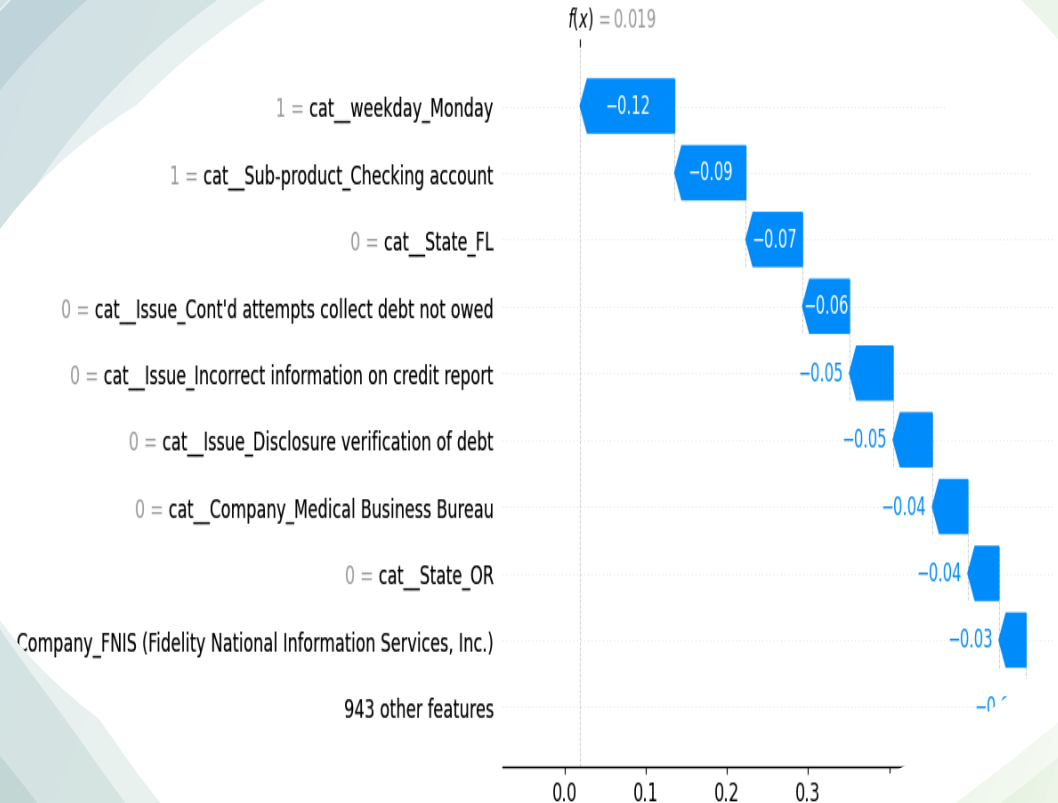


Clase	Precisión	Recall	F1-Score	Soporte
No	0.	0.94	0.90	4990
Yes	0.93	0.84	0.89	4923
Accuracy Total			0.89	9913
ROC AUC			0.86	
Promedio Macro	0.90	0.89	0.89	9913
Promedio Ponderado	0.90	0.89	0.89	9913

# Discusión sobre limitaciones y mejoras

## Dispute

- **Predice mejor los casos negativos** cuando en este caso queremos los positivos
- **No está en pipeline** con preprocesamiento → dificulta implementación.
- **Demasiadas neuronas:** No hay mejoras significativas al usar 64 y 32 neuronas respecto la configuración actual.
- **Estructuras no probadas:** No se han probado otros tipos de estructuras, como las convolucionales.
- **Sobreajuste:** El modelo está sobreajustado, con una precisión de entrenamiento de 0.96 frente a 0.87 en pruebas.
- **Optimización necesaria:** según las necesidades de la empresa.
- **Datos limitados:** Solo se han utilizado 6006 casos, el resto este vacío.
- **Datos desbalanceados:** Solo el 4% de los casos representan situaciones en las que el consumidor no respondió a tiempo, y todos estos casos se disputaron.
- **Baja interpretabilidad:** red neuronal que al usar exclusivamente *onehot encoding* tiene 952 *features* y no esta integrado en pipeline.
- Necesita la respuesta de la compañía, no puede predecir antes





**Preguntas**