Análise exploratoria de bases de datos operativas e xeración de modelos preditivos

Andrea García Pérez

5 de xuño de 2023

Este documento contén un resumo acerca do Traballo de Fin de Mestrado *Análise exploratorio de bases de datos operativos e xeración de modelos preditivos do mercado de traballo* dirixido por don Guillermo López Taboada, Catedrático da Universidade da Coruña e don Carlos Iván Cedrón Alvárez, Risk & Evaluation Senior Analyst en EOS Spain SL [1]. Por motivos de confidencialidade non cabe a posibilidade da publicación da memoria completa.

Resumo

O presente traballo foi desenvolto na área de Reporting & Data Analysis na empresa de EOS Spain SL, unha das principais entidades financeiras especializada no recobro de débedas, área encargada de reportar a información dispoñible nas bases de datos. Por un lado temos o que sería a parte de Reporting propiamente dita que é a encargada da xeración dos cadros de mando para a visualización global dos obxectivos da empresa e poder localizar onde se estarían a producir perdas. Por outra parte podemos atopar a análise de datos, que é na que se centra o marco deste traballo académico. Esta parte do departamento é incipiente, polo que o obxectivo principal da memoria é a elaboración de modelos preditivos que permitan incrementar o número de recuperación de débedas.

O aumento constante da cantidade de datos dispoñibles a nivel interno levou á empresa a crear un área no departamento que se encargase de levar a cado de procesos de limpeza de datos e xeración de modelos preditivos que lle permitisen á empresa optimizar os procesos de recobro e incrementar as cantidades obtidas.

A empresa viña contando con métodos de valoración de carteiras tradicionais, os cales partían do emprego de ferramentas ofimáticas e procesos susceptibles de ser automatizados. A idea que xorde dentro deste marco é poder optimizar todos estes procesos grazas á aplicación de modelos de aprendizaxe estatístico. Sendo isto novidade na compañía comezouse por realizar unha análise exploratoria de datos, elaborando gráficas que se veñan axustando aos coñecementos históricos dos que dispoñemos para saber que o traballo está orientado cara os seus obxectivos.

A popularidade de R e python neste ámbito (ver [2]) de traballou levou a plantexar a súa utilización para analizar a evolución das recuperacións en funcións das diferentes características que poidésemos atopar para os contratos. Para aseguranos en primeiro lugar de que todo está funcionando adecuadamente e de que coñecemos as bases de datos internas, traballamos cunha carteira antiga para ver que os resultados obtidos coinciden co que en realidade se rexistrou.

O primeiro modelo que se plantexou foi un modelo de valoración de débedas xudicializadas (ver [4]). Para levar a cabo dito modelo debemos recurrir ás bases de datos e crear unha táboa de datos coa maior parte de información dispoñible. Unha vez construída a táboa de datos decidimos aplicar modelos de aprendizaxe estatística. Entre os modelos que se deciden levar a cabo, temos que destacar que o obxectivo final é clasificar as nosas débedas en dous grupos: as que pagan e as que non van pagar. É dicir, estamos ante problemas de clasificación supervisada.

Para solventar o problema dispoñemos de modelos como árbores de decisión ou modelos máis complexos como bosques de árbores (onde a unidade individual sabemos que son árbores sinxelos). Grazas ás árbores de decisión sinxelas podemos ver cales son as características dos contratos que teñen unha maior porcentaxe de recobro (daquelas que se atopaban en estado xudicial). Por outro lado, decidimos axustar bosques de árbores, o que se coñece como modelo xgboost.

Cando xeramos un modelo de predición este é entrenado con datos de carteiras antigas e testeado empregado a precisión do modelo (buscamos un modelo o máis preciso posible). Na construción dos modelos da empresa partimos de datos de carteiras antigas. Para a creación do modelo o que facemos é partir da táboa de datos creada, dividindo a mostra nunha parte de entrenamento e noutra de test. Concretamente o entrenamento é xerado co 70% dos datos da mostra e a avaliación lévase a cabo co 30% restante. Ademáis do tamaño das mostras de entrenamento e de test, os modelos necesitan outros parámetros como a profundidade máxima dos erros, o tamaño mínimo dos nodos intermedios, etc. Ditos parámetros poden ser introducidos a man aleatoriamente en función dos nosos coñecementos, pero por sorte contamos con ferramentas adicionais que nos van proporcionar os valores óptimos para o modelo en función do criterio de validación cruzada. Para a selección dos parámetros óptimos empregamos as funcións implementadas en python que nos localizan os parámetros óptimos dentro da rexión de datos que lle indiquemos.

Antes da creación do modelo de contactación propiamente dito, estabamos interesados en levar a cabo unha análise sobre as diferentes accións que se estaban a realizar a nivel interno na empresa para a xestión de débedas. Comezamos así o estudo da seguinte mellor acción (coñecido como Next Best Action, NBA) inspirados en libros como [5]. Grazas a isto conseguimos chegar a facer agrupacións coas accións en función do éxito que tiñan, entendendo por éxito a probabilidade de que posteriormente a algunha das accións do grupo se obtivera unha chamada telefónica onde conseguimos identificar a persoa e establecer un acordo para o pago das débedas. Para o grupo con maior porcentaxe de contactabilidade inmediata descubrimos que o ideal sería realizalas o día 30 do mes ás 13h, sempre e cando este cadrase nun mércores e non fose víspera de festivo. A maiores da situación ideal, xeramos unha árbore iterativa que en función do día no que nos atopasemos nos devolvese as mellores horas para realizar a acción.

Tralo bo funcionamento do primeiro modelo, decidiuse implementar técnicas de aprendizaxe estatística en máis estudos, entre eles identificar aqueles contratos para os cales é posible que consigamos contactar de maneira directa, ao menos a primeira vez, durante o primeiro ano de xestión interna dende a compra da carteira. Neste caso o modelo será similar ao anterior, grazas ás características das débedas clasificamos en contacto no primeiro ano si e non. De novo, ante este problema de clasificación aplicamos modelos xgboost e incluso probamos a executar un modelo de regresión binaria, vendo que funciona mellor o bosque de árbores.

Cabe destacar que grazas aos modelos xgboost podemos obter o grupo ao que pertencerán os novos contratos (pagan ou non; contacto ou non) pero, a maiores do anterior, podemos obter a probabilidade de que pertenza a cada un dos grupos. Para a implementación dos modelos, primeiramente consideramos un axuste dos parámetros manuais obtendo unha precisión de case un 72%. Se axustamos os modelos en función da optimización dos parámetros non conseguimos aumentar a precisión do modelo. Vexamos o Cadro 1 no que observamos a comparación dos resultados dos modelos en función dos parámetros escollidos.

	Semilla de aleatoriedade	Radio de aprendizaxe	Número de estimadores	Máxima profundidade	Precisión
Parámetros aleatorios	42	0.1	600	3	0.71247230
Parámetros óptimos	Non empregado	Non empregado	100	20	0.71192973

Cadro 1. Comparación dos resultados do modelo xgboost de contactación según parámetros

No modelo de contactación tamén estaban interesados en estimar, o número de días ata que se produce o primeiro contacto directo con un determinado contrato. Para iso, neste caso a variable a predicir é a cantidade de días polo que imos aplicar un modelo de regresión lineal. Obtemos un erro de máis ou menos 90 días na estimación.

Para visualizar o que se estaba a levar a cabo, cando se produxo a compra dunha carteira nova, decidimos aplicar ambos modelos xerados e obter as predicións de recobro. Para darlle visualización as tarefas do departamento, creouse un informe no que se almacenasen as recuperacións reais que se estaban a obter da carteira nova e ademais as que predicía o modelo. Nun diagrama de liñas tiñamos a comparación por meses acumulados dende a data de compra da carteira e podiamos observar que de vez en cando se ían solapando.

Á par dos resultados obtidos e grazas a multitude de funcionalidades de python que fomos descubrindo, os obxectivos e tarefas a abarcar polo departamento foron en aumento. Entre eles apareceu unha nova tarefa, enriquecer as bases de datos. Para levar a cabo isto, traballamos coa libraría request de python [3], coa cal solicitabamos o acceso a unha determinada URL empregando unha clave API. Grazas a isto, conseguimos acceder a unha plataforma externa, mailgun, a cal nos proporcinaba o estado no que se atopaba unha dirección de enderezo electrónico, o estado da mensaxe enviada e tamén do buzón do correo electrónico. Grazas a isto, a empresa pode deixar de enviar correos electrónicos a direccións que non existen e van devolver continuamente erros.

A oportunidade de realizar este TFM de modalidade B na compañía EOS Spain SL, desenvolto na súa sede da Coruña, foi enriquecedora e moi positiva, permitindo aplicar (e ampliar notariamente) os meus coñecementos sobre a parte financeira e a aplicación estatística. Trala finalización da execución deste TFM, puiden continuar formando parte da compañía e ampliando os meus coñecementos.

Referencias

[1] Enlace EOS Spain SL: https://es.eos-solutions.com. Última data de acceso: 05-06-2023.

[2] Machine Learning con python ou R https://www.cienciadedatos.net. Última data de acceso: 05-06-2023

[3] Libraría python para a extracción de datos de páxinas web https://pypi.org/project/requests/. Última data de acceso: 05-06-2023

[4] Casos de éxito de Nexus Integra https://nexusintegra.io/es/casos-de-exito/. Última data de acceso: 05-06-2023

[5] Roy, D., Fernando, B. (2022). Predicting the Next Action by Modeling the Abstract Goal. arXiv preprint arXiv:2209.05044. Última data de acceso: 05-06-2023