Informe de seguimiento

Miembros del grupo:
- Ricardo Ocaña
- Miguel Sempere
- Álvaro Rodríguez
- Álvaro Ferro
- Jorge Casañ
- Luis Llera

En el presente documento se realiza un informe de seguimiento. Este informe pretende realizar un seguimiento de nuestra variable objetivo: "Tasa de Mora". Para ello, se mostrará cómo evoluciona la predicción en función de los diferentes modelos y cómo, con algunas mejoras propuestas, la predicción mejora.

A modo introductorio, conviene recordar que las variables utilizadas definitivamente para el modelo, antes de realizar mejoras, son las siguientes:

```
['ListingCategory (numeric)',
 'Occupation',
 'EmploymentStatus',
 'EmploymentStatusDuration',
 'IsBorrowerHomeowner',
 'CreditScoreRangeLower',
 'CreditScoreRangeUpper',
 'CurrentCreditLines',
 'OpenCreditLines',
 'TotalCreditLinespast7years',
 'OpenRevolvingAccounts',
 'OpenRevolvingMonthlyPayment',
 'InquiriesLast6Months',
 'TotalInquiries',
 'CurrentDelinquencies',
 'AmountDelinquent',
 'DelinquenciesLast7Years',
 'TotalTrades',
 'TradesNeverDelinquent (percentage)',
 'IncomeRange',
 'StatedMonthlyIncome',
 'ScorexChangeAtTimeOfListing',
 'LoanOriginalAmount',
 'PercentFunded',
 'LoanFirstDefaultedCycleNumberQ']
```

Antes de realizar los modelos predictivos, se realiza un análisis de estabilidad de las variables. Esto está explicado detalladamente en la práctica 3; sin embargo y a modo resumen, cabe recordar que el análisis de estabilidad pretende, como su nombre dice, estudiar la estabilidad de las variables y la significatividad de los cambios. Existen umbrales que miden la estabilidad. A continuación, se muestra dicho análisis para las variables escogidas para nuestro modelo:

	feature	PSI
4	IsBorrowerHomeowner	0.000000e+00
23	PercentFunded	2.141008e-08
15	AmountDelinquent	7.428038e-05
14	CurrentDelinquencies	3.430635e-04
16	DelinquenciesLast7Years	1.055222e-03
13	TotalInquiries	3.452111e-03
12	InquiriesLast6Months	4.000090e-03
19	IncomeRange	9.661644e-03
18	TradesNeverDelinquent (percentage)	9.802307e-03
20	StatedMonthlyIncome	1.143998e-02
9	TotalCreditLinespast7years	1.401206e-02
17	TotalTrades	1.462368e-02
1	Occupation	1.775389e-02
7	CurrentCreditLines	1.799375e-02
10	OpenRevolvingAccounts	2.199308e-02
8	OpenCreditLines	2.774108e-02
21	ScorexChangeAtTimeOfListing	2.969534e-02
6	CreditScoreRangeUpper	3.220120e-02
5	CreditScoreRangeLower	3.220120e-02
3	EmploymentStatusDuration	3.250906e-02
11	OpenRevolvingMonthlyPayment	4.459233e-02
0	ListingCategory (numeric)	2.378023e-01
22	LoanOriginalAmount	2.832083e-01
24	LoanFirstDefaultedCycleNumberQ	4.391166e-01
2	EmploymentStatus	5.479469e-01

Solo las superiores a 0,25 suponen cambios significativos. Todo lo que sea por debajo de dicho umbral, pero por encima de 0,10; es mejor estudiar de forma personalizada. Para todo lo demás, no es necesario.

A continuación, se va a mostrar, en base a las variables escogidas y al análisis de estabilidad anterior, el resultado de los diferentes modelos predictivos realizados.

<u>Para el modelo de regresión logística, el resultado de la predicción es:</u>

```
1 get_auc(y_tr, pred_tr), get_auc(y_val, pred_val), get_auc(y_oot, pred_oot)
```

```
(0.6838508477949512, 0.6671933928665152, 0.6722968401412276)
```

Para el modelo de random forest, el resultado de la predicción es:

```
14 # Get AUC metrics in all subsets
15 get_auc(y_tr, pred_tr), get_auc(y_val, pred_val), get_auc(y_oot, pred_oot)
```

```
(0.7103581967727758, 0.6771205884337675, 0.6845644448176392)
```

Para el modelo de XGBoost, el resultado de la predicción es:

```
6 get_auc(y_tr, pred_tr), get_auc(y_val, pred_val), get_auc(y_oot, pred_oot)
```

```
(0.7123293363233906, 0.6801148206881555, 0.6809619398105502)
```

A simple vista se ve como el modelo de regresión logística es el peor de los tres y como el modelo de random forest y el de XGBoost son muy parecidos, arrojando resultados sobre el conjunto de validación muy buenos. Esto confirma que nuestros modelos son buenos.

Ahora es interesante comprobar cómo, a través de Reject Inference y a través de Feature Engineering podemos obtener resultados diferentes y qué se ha hecho para ello.

Gracias a Reject Inference y cómo se puede comprobar en el notebook específico, con las mismas variables conseguimos que el modelo hecho con regresión logística nos arroje un resultado superior al anterior:

```
1 get_auc(y_tr_new, pred_oot)
```

```
0.7907963543280936
```

Por último, gracias a la ingeniería de variables, se modifican algunas, se realiza su correspondiente análisis de estabilidad y se vuelve a realizar una predicción de la tasa de mora. Sin embargo, el resultado es peor, por lo que nos quedamos con las variables escogidas desde un principio. A continuación, se muestra el score final, pero el análisis psi del modelo tras la ingeniería de variables se encuentra en su correspondiente notebook de la práctica:

```
1 get_auc(y_tr, pred_tr), get_auc(y_oot, pred_oot)
```

```
(0.6230307558897848, 0.6068645999550861)
```