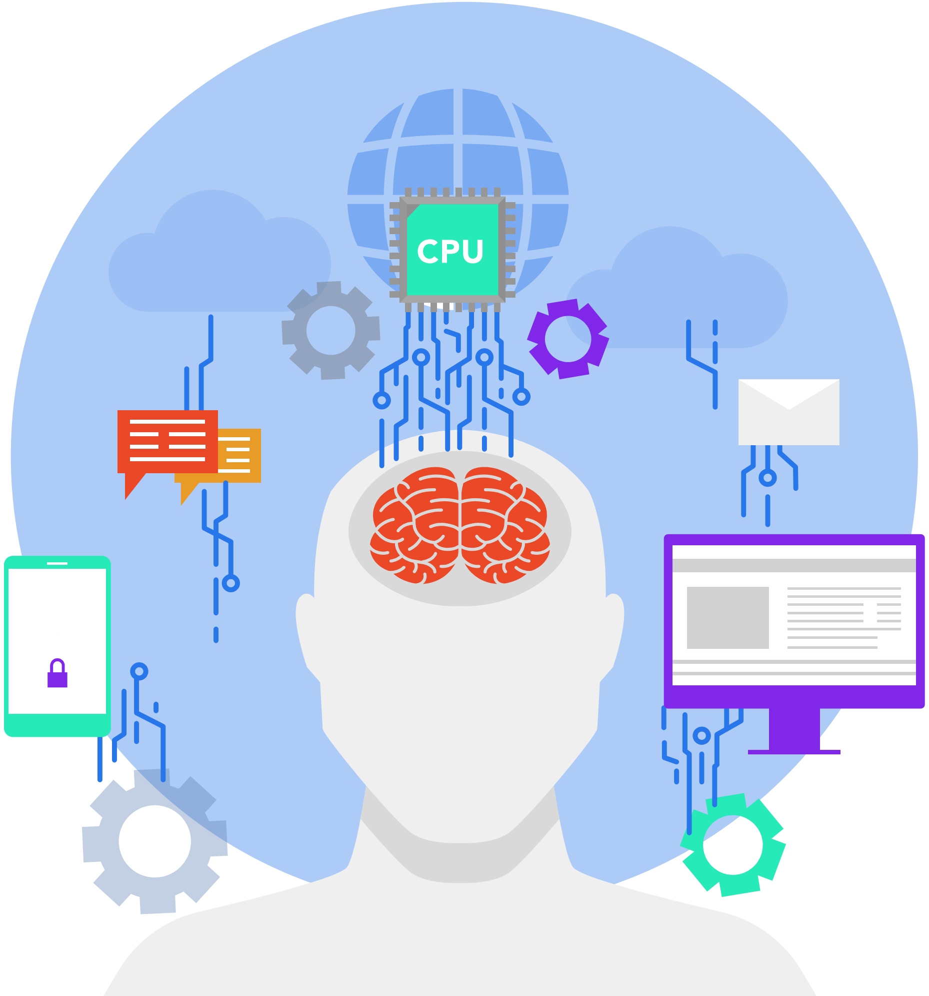
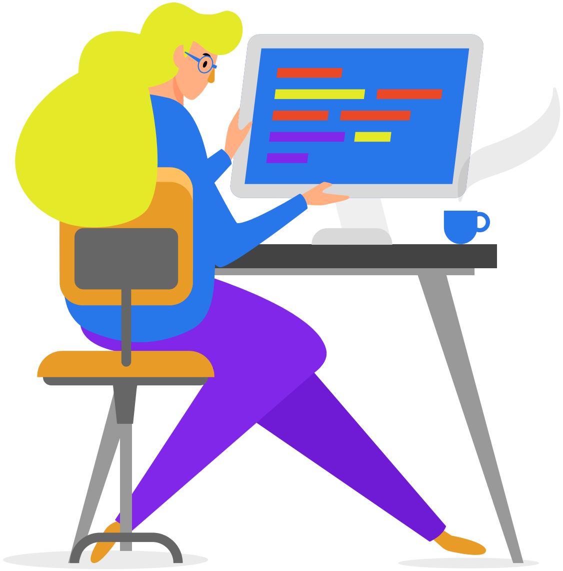


New product analysis for Xepelin

Cristhian Plazas Ortega
Científico de Datos

Route map of the Work



01 EDA analysis

02 Hypothesis Tests

03 Modelling the baseline

04 Improving the modelling stage

05 Forecasting status

06 Future Analysis and Conclusions

EDA Analysis

1200 Filas

7 Columnas

DataTypes	Frequency	Columns
Float64	2	amount, amountfinancedbyXepelin
Int64	3	PayerId, ReceiverId, InvoicedId
Datetime64[ns]	1	paidAt
Object	1	status

Column	Missing %
PayerId	0
ReceiverId	0
invoiceId	0
paidAt	26
amount	0
amountfinancedByXepelin	0
status	0

Status Column with Missing paidAt	Missing %
PROCESSING	100
FAILED	100

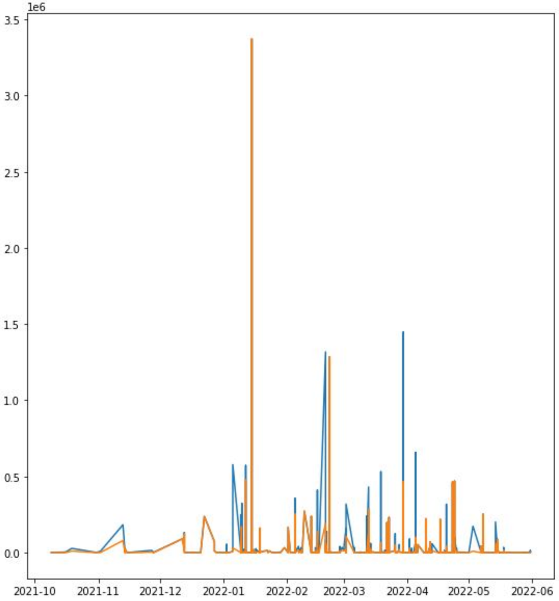
EDA Analysis

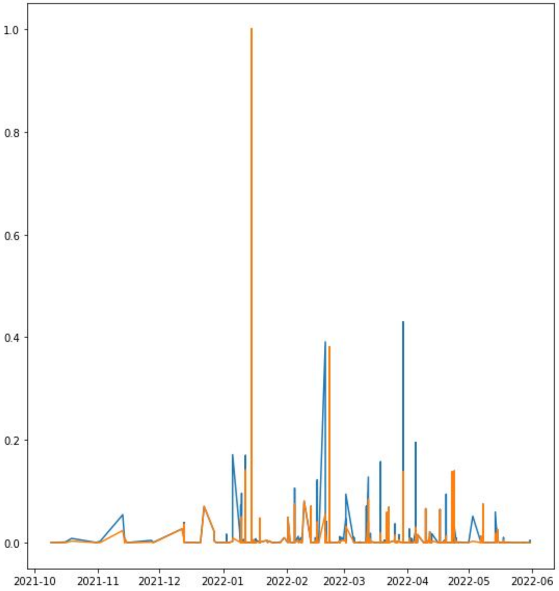
Unity-based normalization for amount and amountfinancedByXepelin features

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Blue: amount

Orange: amountfinancedByXepelin



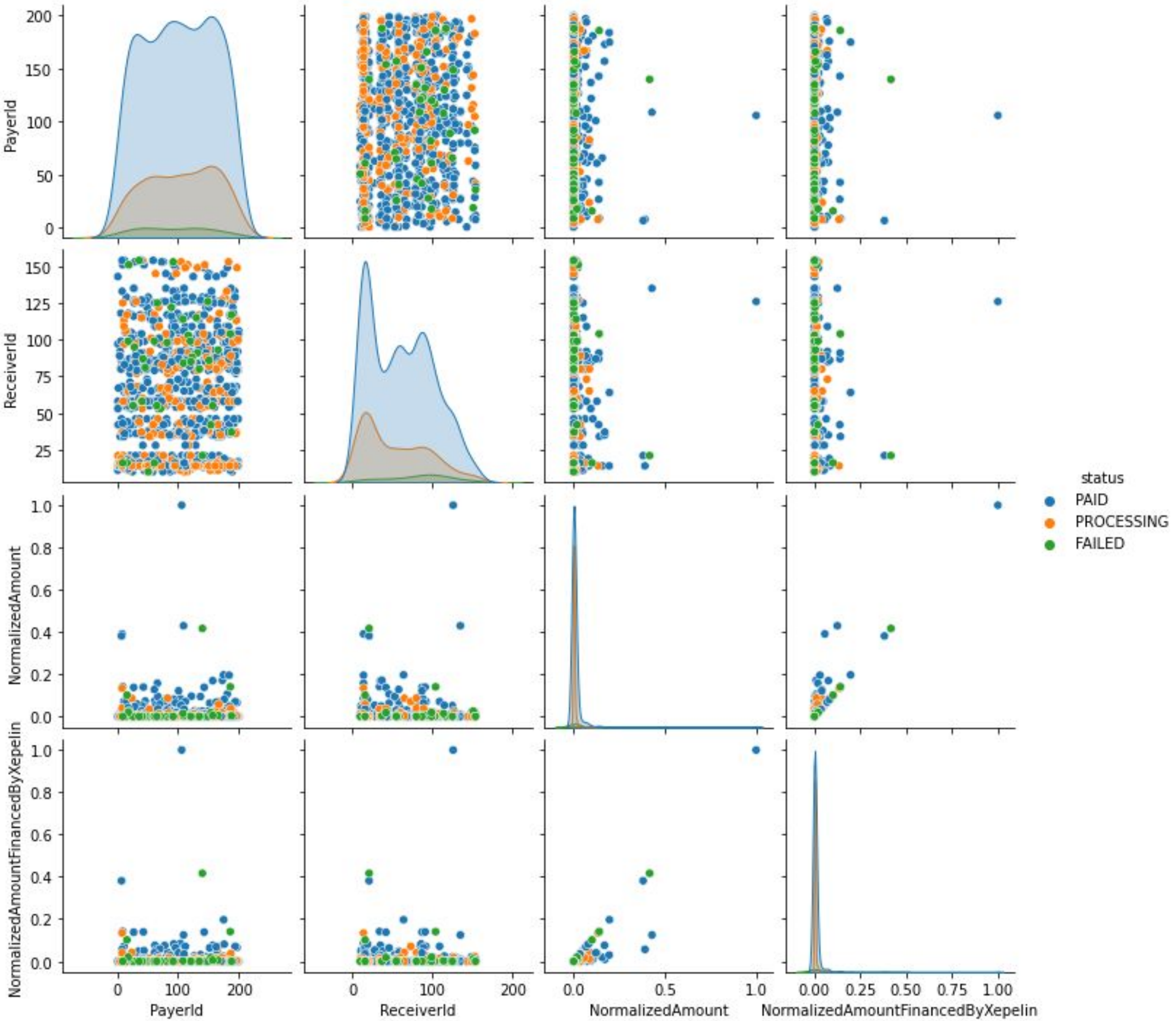


EDA Analysis

- Anomalies Data seems to be present on amount fields
- PayerId seems a normal distribution
- ReceiverId and both amount fields seems a asymetric distribution
- Amount fields seems keeping a relation with status variable

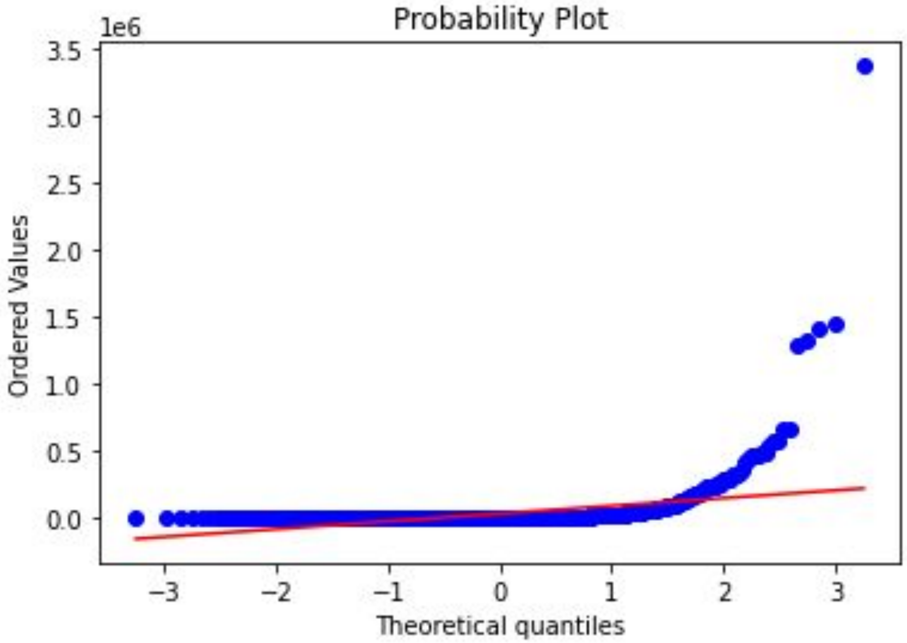
Questions

- Why do payments fail?
- Is there a relationship among status and amount fields?
- Is there a relationship among status and payers fields



Hypothesis

Amount field may be directly correlated with status field. Because of amount keeps an asymmetric distribution the dependence relation calculated used the median equality theory.



Shapiro Test

0.0

P-Value

Asymmetric Distribution

Hypothesis

Amount seems keep a weak inverse correlation with status. High amounts frequently were paid against low amounts which failed the payment.

Median Equality Test

0.047

P-Value

Spearman

-0.10

Correlation

Amount and Status fields may keep a dependence

Hypothesis

Amount doesn't keep a relation with Payer. If it existed which users can signify a change of status would be ideal.

Chi2 Test

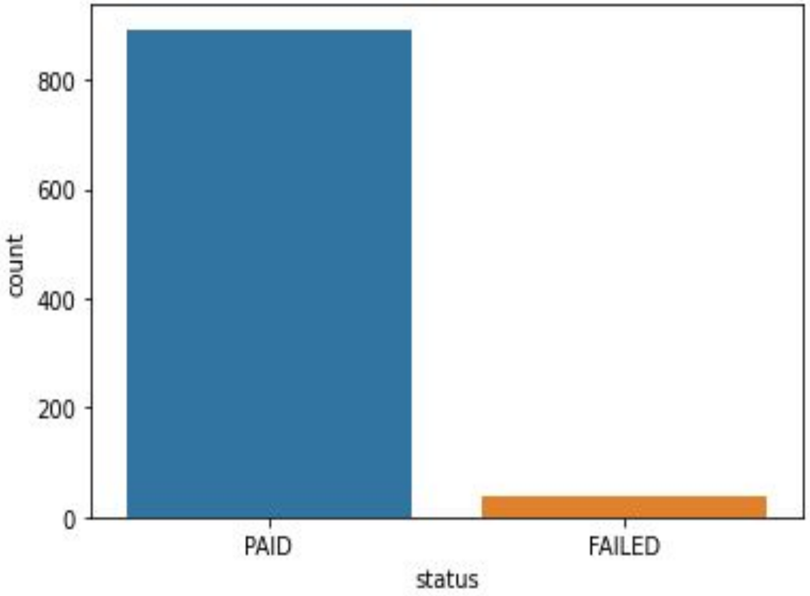
0.99

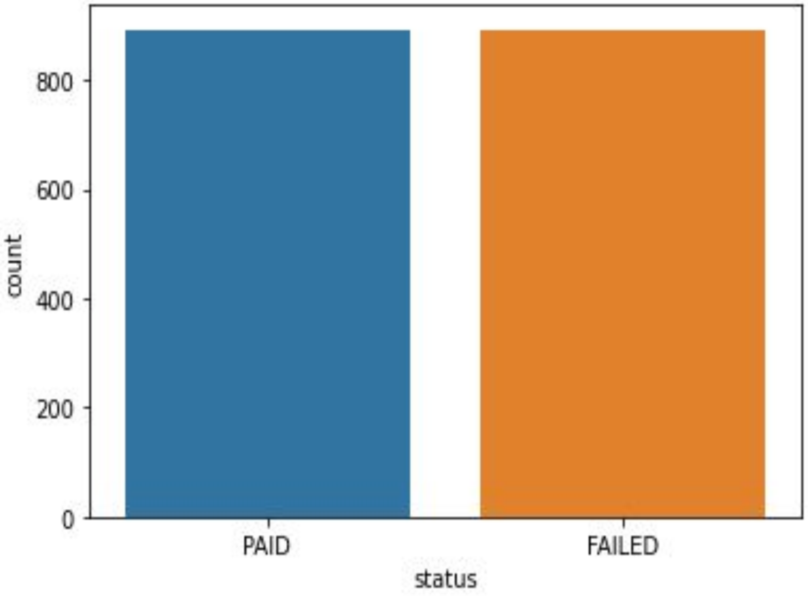
P-Value

Amount and Payer fields may not keep a dependence

Modelling BaseLine

Status is the target field. Forecast if a Preprocessing status may be a Paid or Fail one may help to understand how many is going to be paid and financed by Xepeling. Target is imbalanced but using SMOTENC (Nominal/Continue) method it will being balanced.





Modelling BaseLine

Some basics classifiers like RandomForest, LGBM and Logistic Regression, were used to modelling the data. The best classifier is improved using GridSearchCV method.

	RF Precision	LGBM Precision	LR Precision
FAILED	0.88	0.93	0.60
PAID	0.96	0.98	0.58
accuracy	0.92	0.95	0.59

	RF Recall	LGBM recall	LR Recall
FAILED	0.95	0.99	0.65
PAID	0.88	0.92	0.56
accuracy	0.92	0.95	0.60

Confusion matrix provides a worthy information. LGBM classifier seems give a better performance againts the other ones. However, false negative indicator needs to be improved

	True Values	
Predictive Values	220	3
	16	207

Improving the modelling stage

LGBM Model passed through GridSearchCV optimization.

GridSearch Optimization seems not improve a lot the model. However, only some hyperparameters were optimized. Time of modelling is still keeping a good performance for what optimizing the rest of the hyperparameters may provide better results.

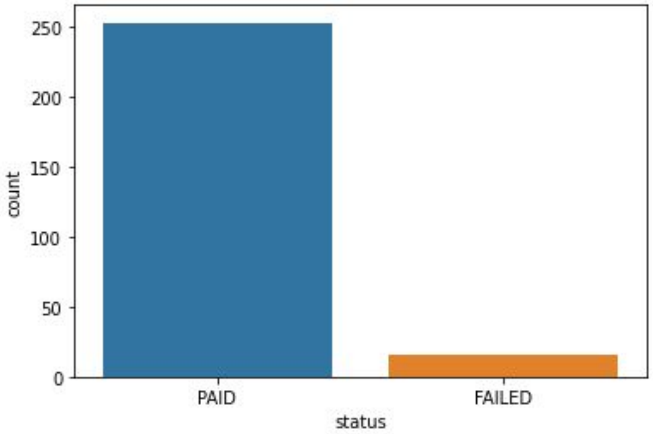
	GS Precision	GS Recall
FAILED	0.93	0.98
PAID	0.98	0.93
accuracy	0.95	0.95

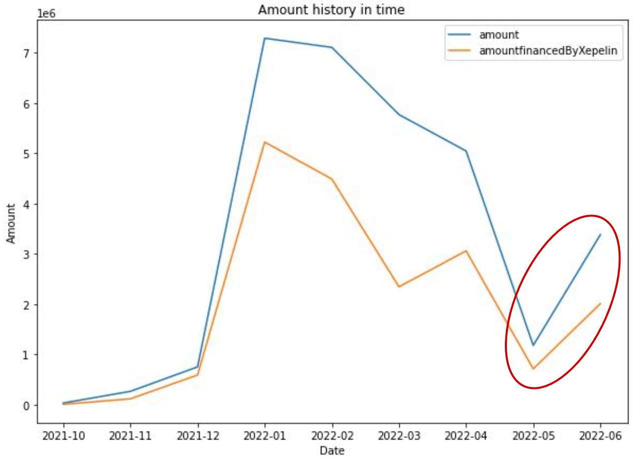
17m54.6s

Time of performance

	True Values	
Predictive Values	220	3
	15	207

Forecasting





3380613.4

Amount to Pay on June 2022

2011202.9

Amount financed by Xepelin on June 2022

Future Analysis and Conclusions

- Althoguh the statement of the problem did not describe whether status should be used as the target variable, this assumption was assumed as a solution to the exercise through a classifier. Given the above, other solutions could be taken into account, such as the development of a time series based-model.
- Generally small datasets has Bias, Overfit and Outliers problems.
  - Simpler models are more suitable to implement. RandomForest and LogisticRegression are part of these simpler models but LGBM gave better results.
  - Regularization prevents the overfitted models
- More features could be added.
  - Type of the industry of the payer
  - Kind of the product or service to pay
  - Etc.
- False Negative is a big problem. Overestimate the payment is always better than underestimate it. Even so, 15 FN represents only the 0.0125% of the trained model.