ML and Dashboard final Project

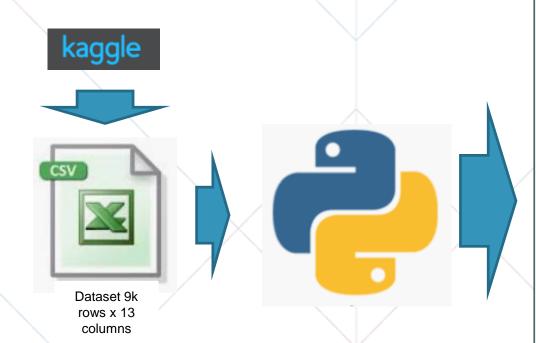
Predicting whether the borrower will pay back their loan in full

AGENDA



- 1. Process
- 2. Model
- 3. Exploratory
- 4. Dashboard
- 5. Takeaways

Machine Learning and ETL process - Dash



- Exploratory Analysis
 - Pandas
 - NumPy
 - Seaborn
 - Matplotlib
- 2. Dashboard Dash.
 - Flask
 - Html
 - CSS
 - Plotly
- 3. Regression Model Logistic
- 4. Machine Learning
 - Decision Tree
 - Random Forest
 - Grid Search

Exploratory analysis

- 0.3

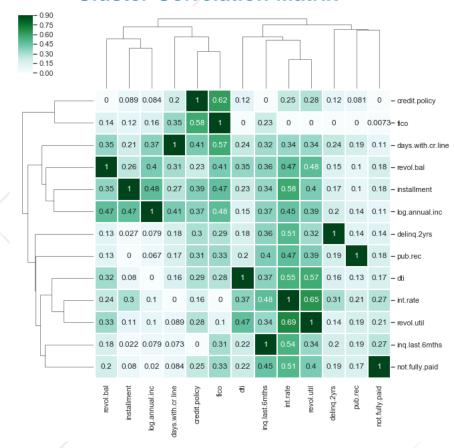
- -0.3

--0.6

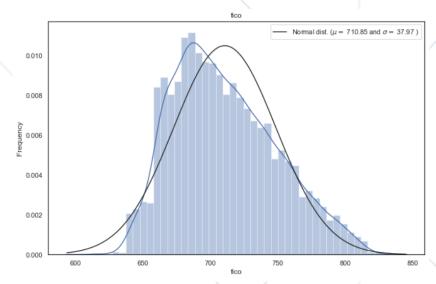
Correlation Matrix

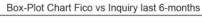
credit.policy	1	-0.29	0.059	0.035	-0.091	0.35	0.099	-0.19	-0.1	-0.54	-0.076	-0.054	-0.16
int.rate	-0.29	1	0.28	0.056	0.22	-0.71	-0.12	0.093	0.46	0.2	0.16	0.098	0.16
installment	0.059	0.28	1	0.45	0.05	0.086	0.18	0.23	0.081	-0.01	-0.0044	-0.033	0.05
log.annual.inc	0.035	0.056	0.45	1	-0.054	0.11	0.34	0.37	0.055	0.029	0.029	0.017	-0.033
dti	-0.091	0.22	0.05	-0.054	1	-0.24	0.06	0.19	0.34	0.029	-0.022	0.0062	0.037
fico	0.35	-0.71	0.086	0.11	-0.24	1	0.26	-0.016	-0.54	-0.19	-0.22	-0.15	-0.15
days.with.cr.line	0.099	-0.12	0.18	0.34	0.06	0.26	1	0.23	-0.024	-0.042	0.081	0.072	-0.029
revol.bal	-0.19	0.093	0.23	0.37	0.19	-0.016	0.23	1	0.2	0.022	-0.033	-0.031	0.054
revol.util	-0.1	0.46	0.081	0.055	0.34	-0.54	-0.024	0.2	1	-0.014	-0.043	0.067	0.082
inq.last.6mths	-0.54	0.2	-0.01	0.029	0.029	-0.19	-0.042	0.022	-0.014	1	0.021	0.073	0.15
delinq.2yrs	-0.076	0.16	-0.0044	0.029	-0.022	-0.22	0.081	-0.033	-0.043	0.021	1	0.0092	0.0089
pub.rec	-0.054	0.098	-0.033	0.017	0.0062	-0.15	0.072	-0.031	0.067	0.073	0.0092	1	0.049
not.fully.paid	-0.16	0.16	0.05	-0.033	0.037	-0.15	-0.029	0.054	0.082	0.15	0.0089	0.049	1
	aredit.policy	int.rate	installment	log.annual.inc	ŧ	fico	days.with.cr.line	revol.bal	revol.util	ing.last.6mths	deling.2yrs	pub.rec	not.fully.paid

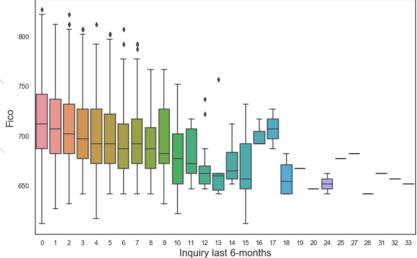
Cluster Correlation Matrix



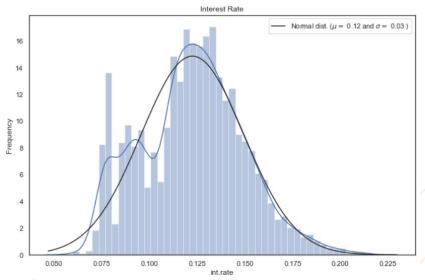
Fico



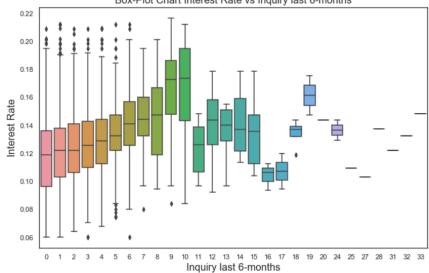




Interest Rate



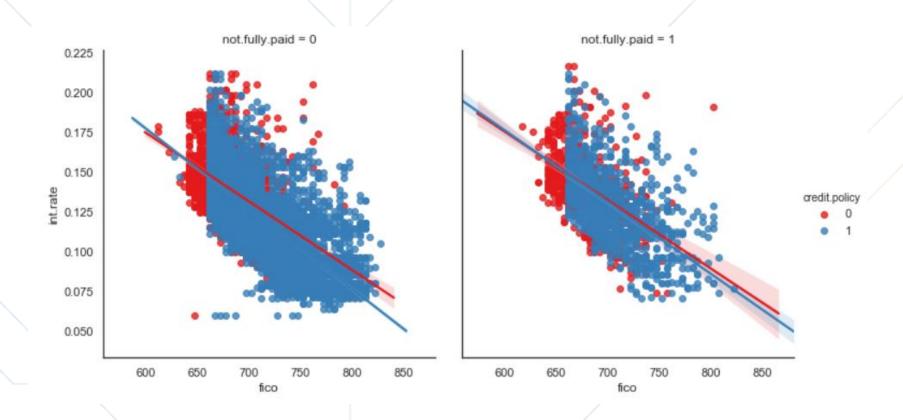
Box-Plot Chart Interest Rate vs Inquiry last 6-months



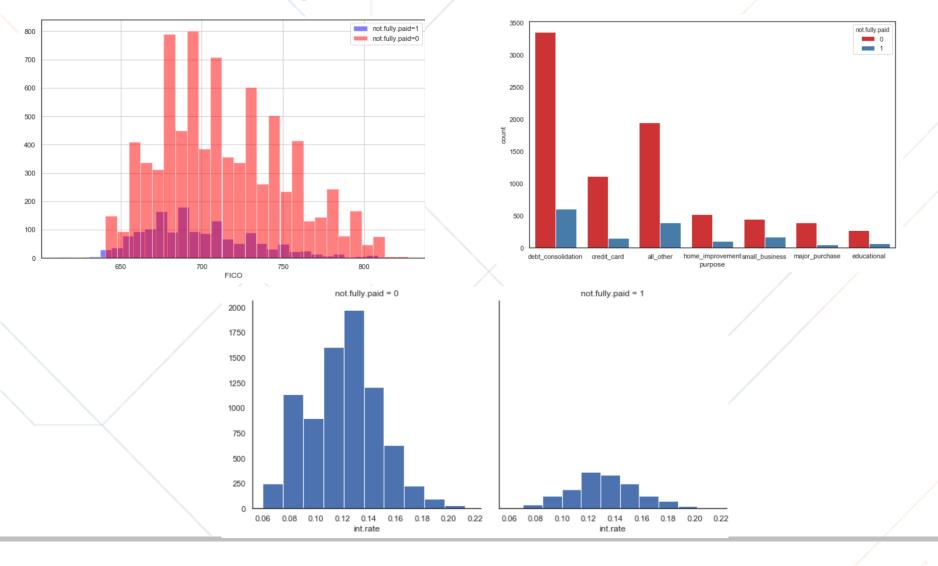


Dash: Dashboard

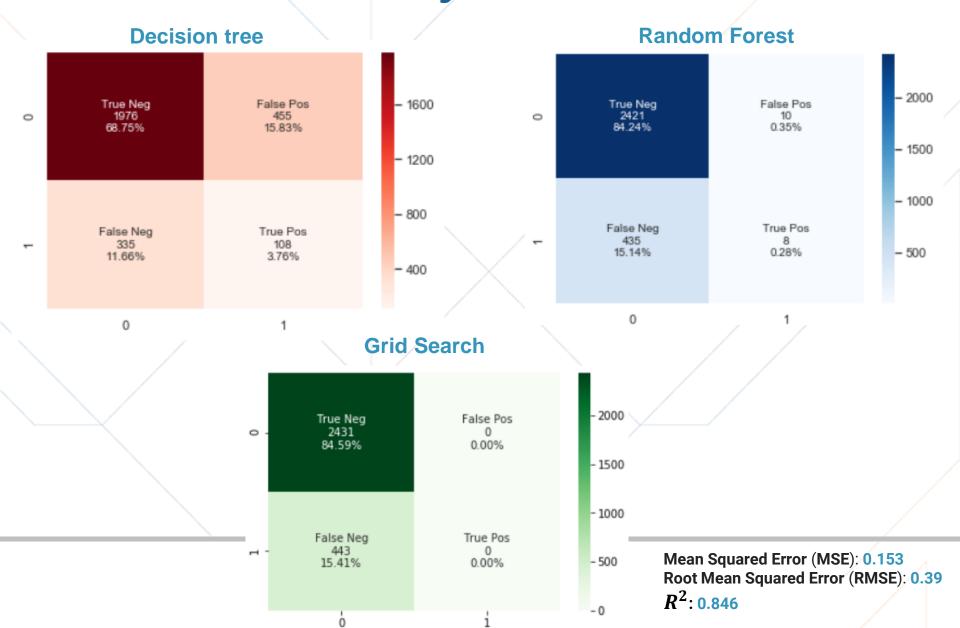
Separating the effect of not fully paid



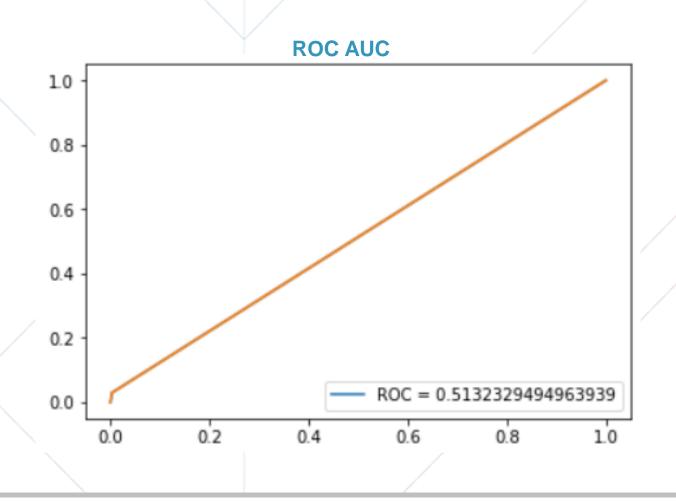
Separating the effect of not fully paid



Which model do you think is better?



ROC: Receiver operating characteristic



Conclusion: pay attention to lonely numbers

In order to create a logistic regression model we must consider the following:

- Data Cleansing (50% of the time)
- Exploring (20%)
- Define the model (20%)
- Ask the right question (5%)
- Rethink your model (5%)
- ROC for Logistic Regression with categorical features (0 to 1)
- Add more features (the model is barely better than random)

Thank Vou

Appendix