ML and Dashboard final Project

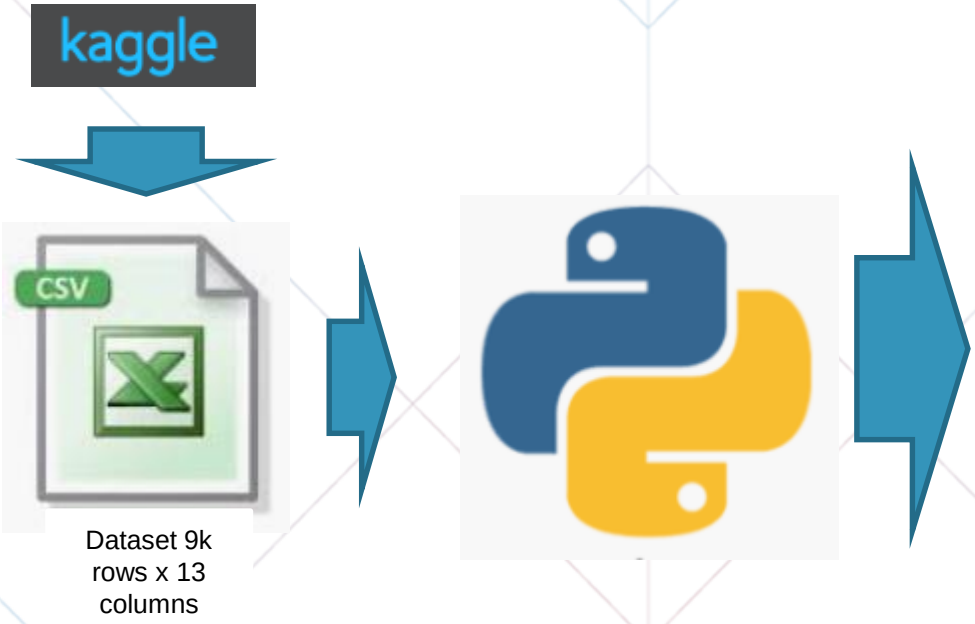
Predicting whether the borrower will pay back their loan in full

Guillermo Sáez

AGENDA

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

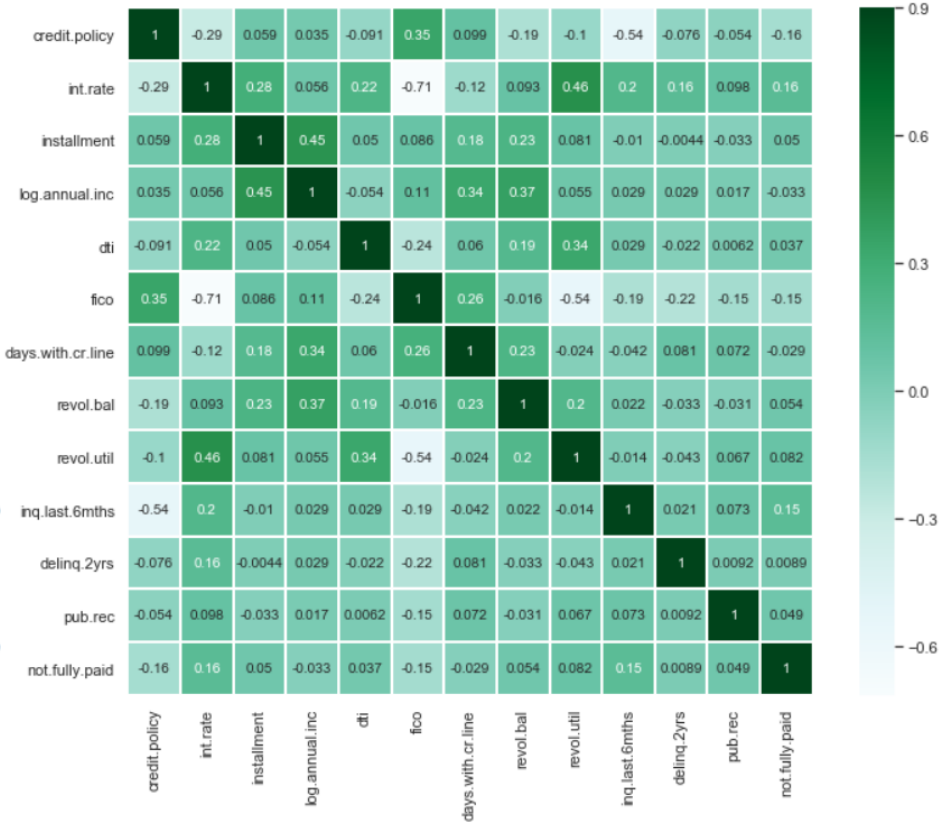
Machine Learning and ETL process - Dash



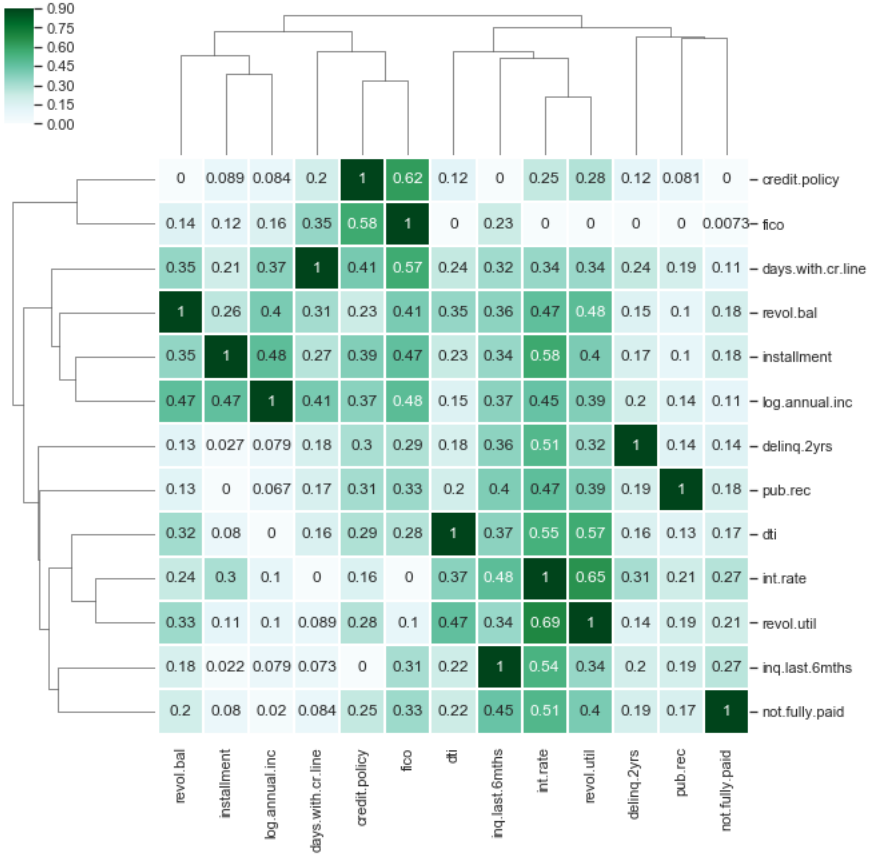
1. Exploratory Analysis
   - Pandas
   - NumPy
   - Seaborn
   - Matplotlib
2. Dashboard – Dash.
   - Flask
   - Html
   - CSS
   - Plotly
3. Regression Model – Logistic
4. Machine Learning
   - Decision Tree
   - Random Forest
   - Grid Search

Exploratory analysis

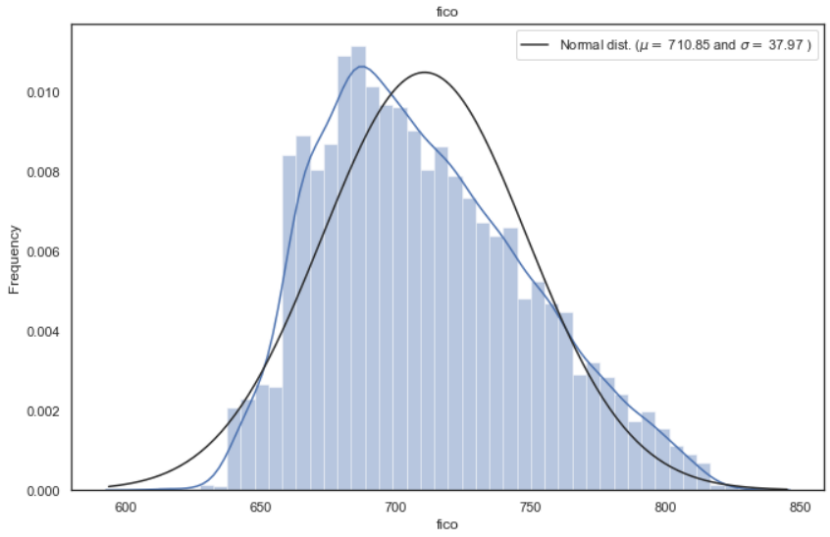
Correlation Matrix

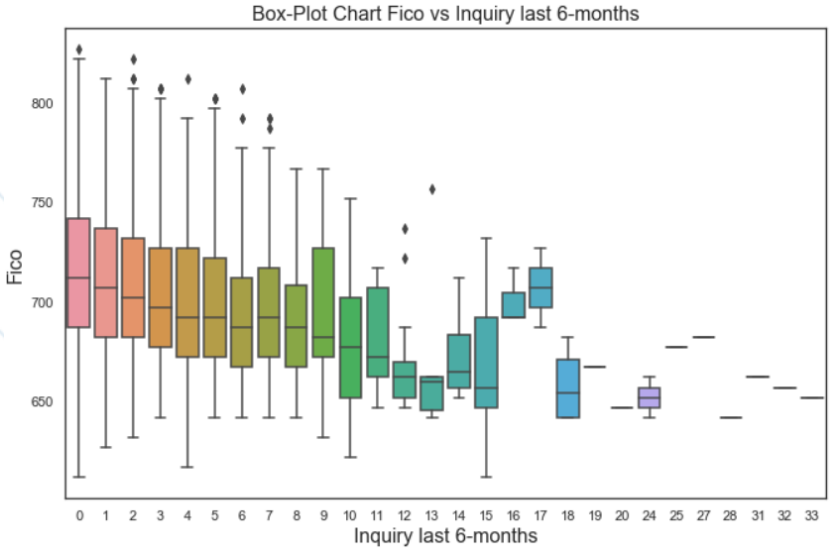


Cluster Correlation Matrix

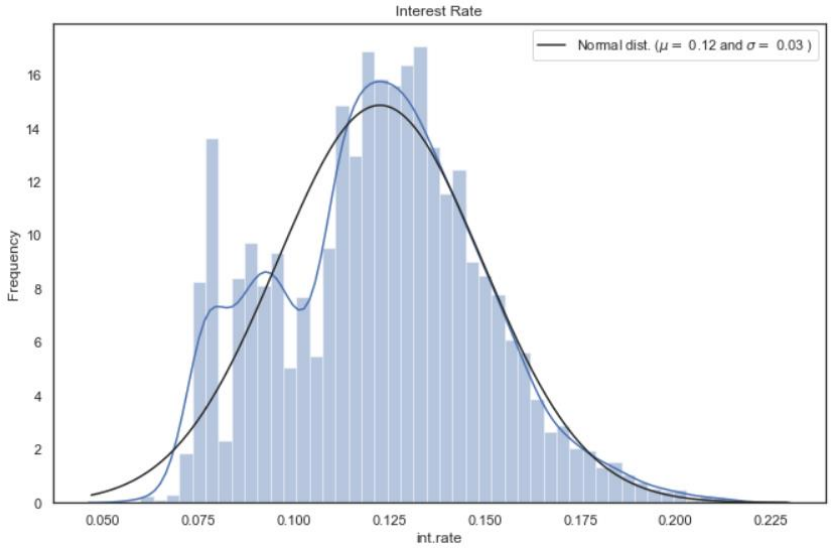


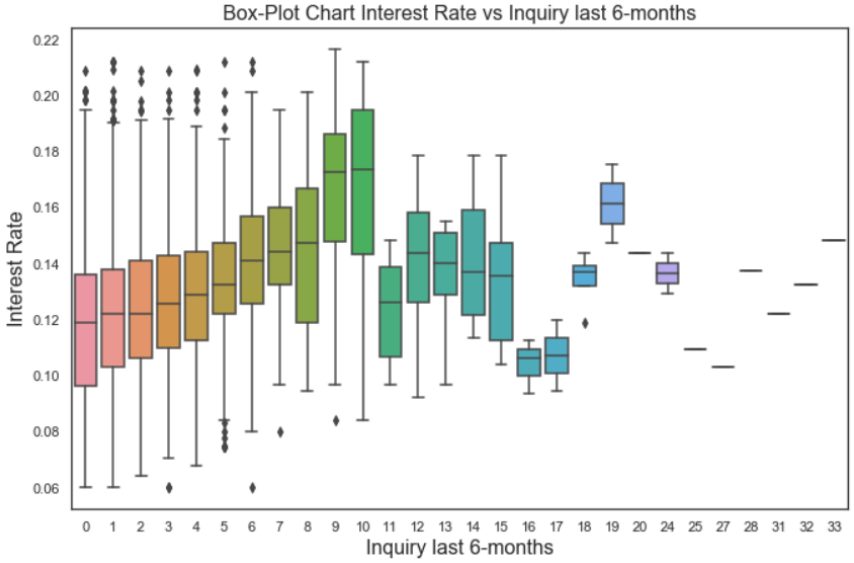
Fico



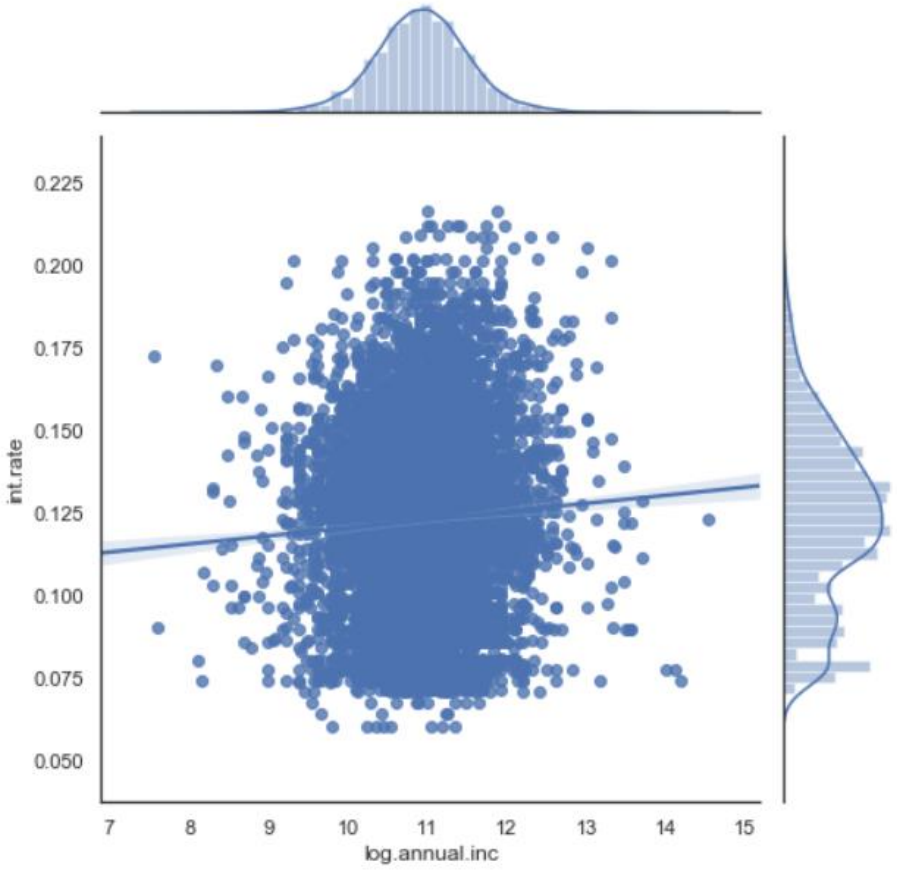


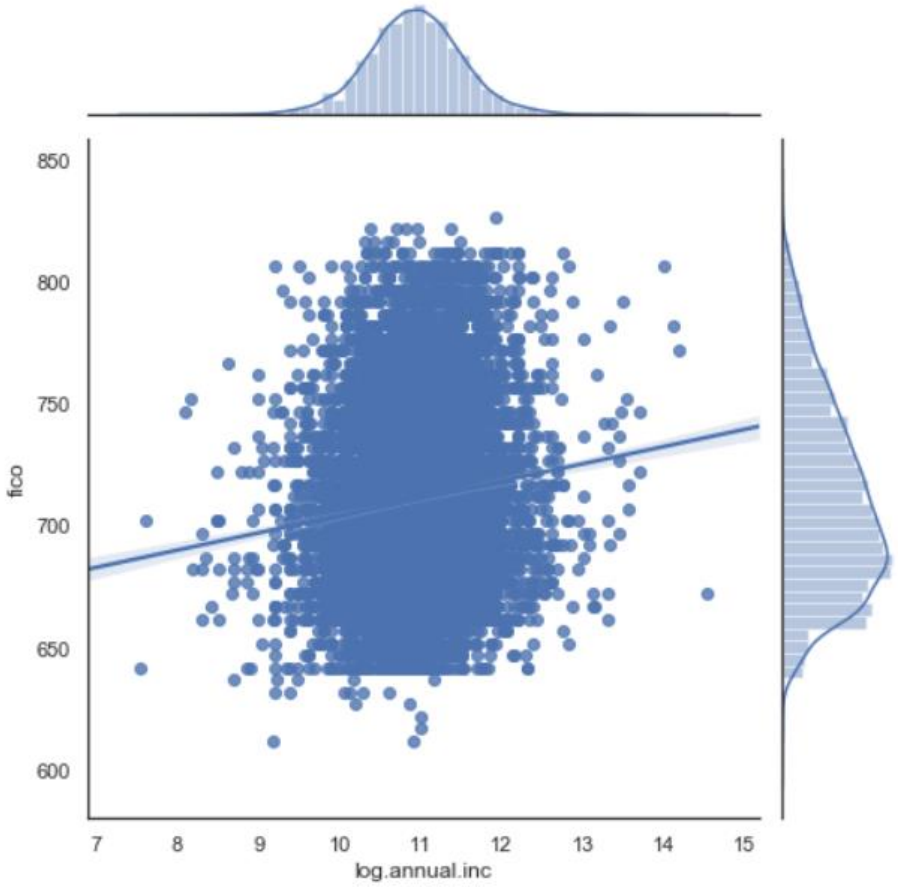
Interest Rate





Sales vs Fico and Interest rates

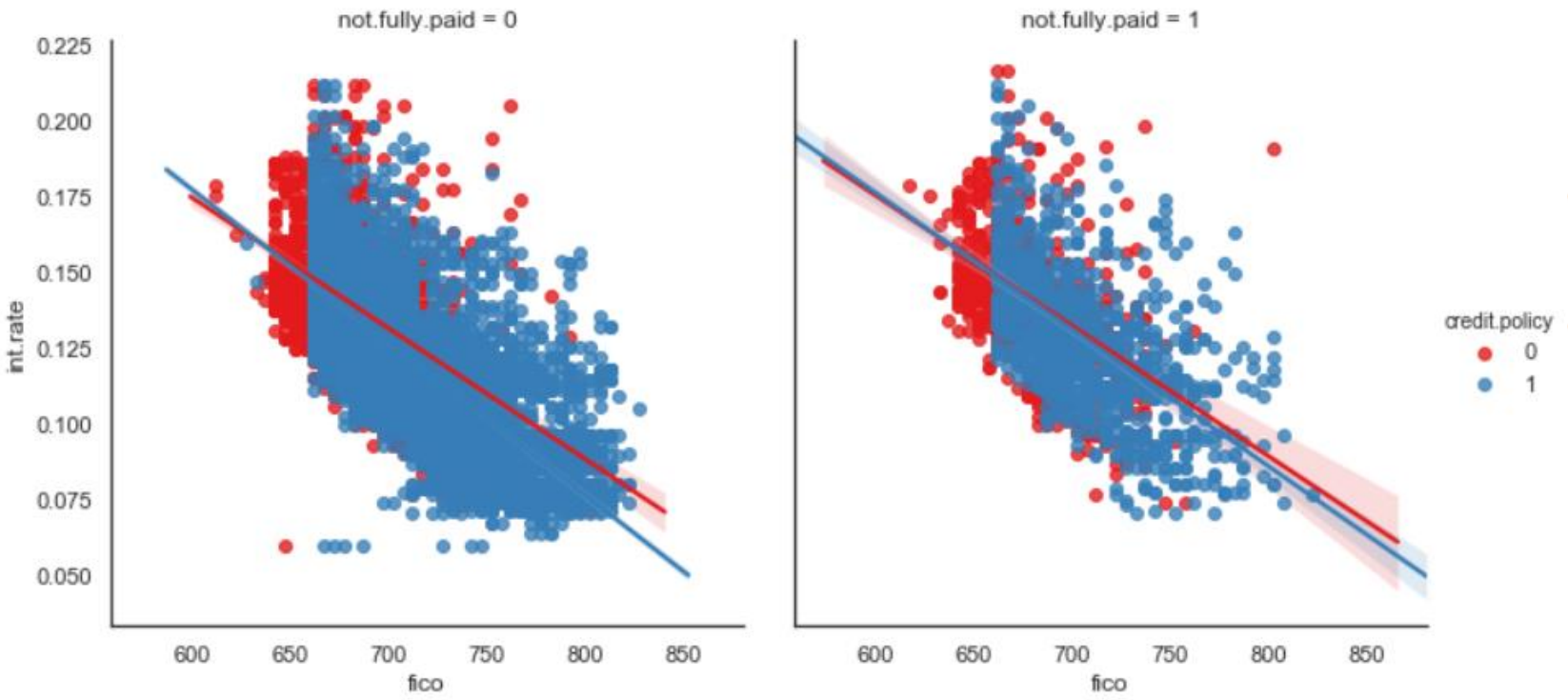




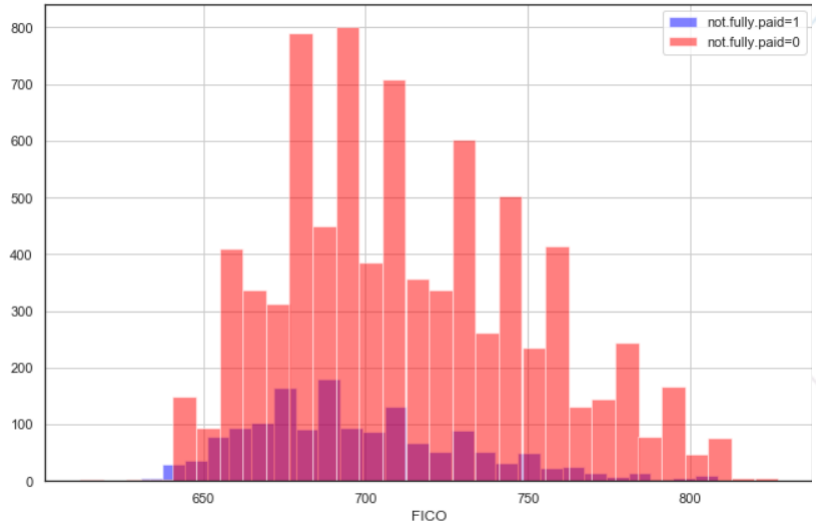
Dash: Dashboard

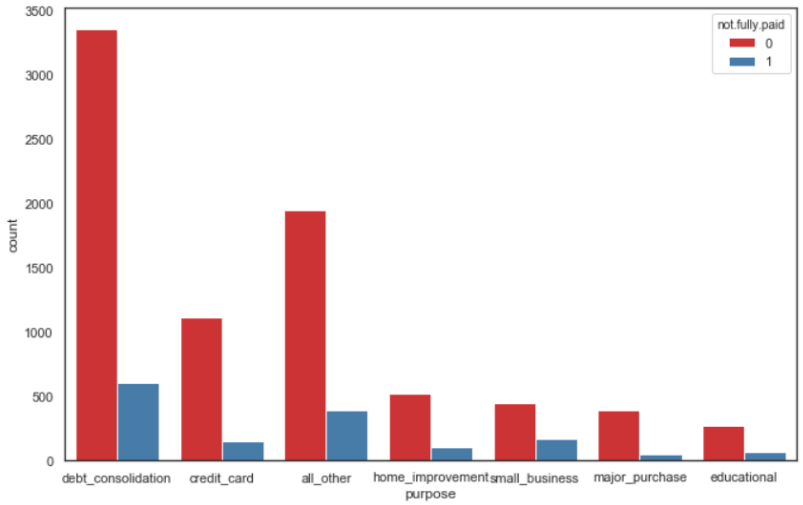
Guillermo Sáez

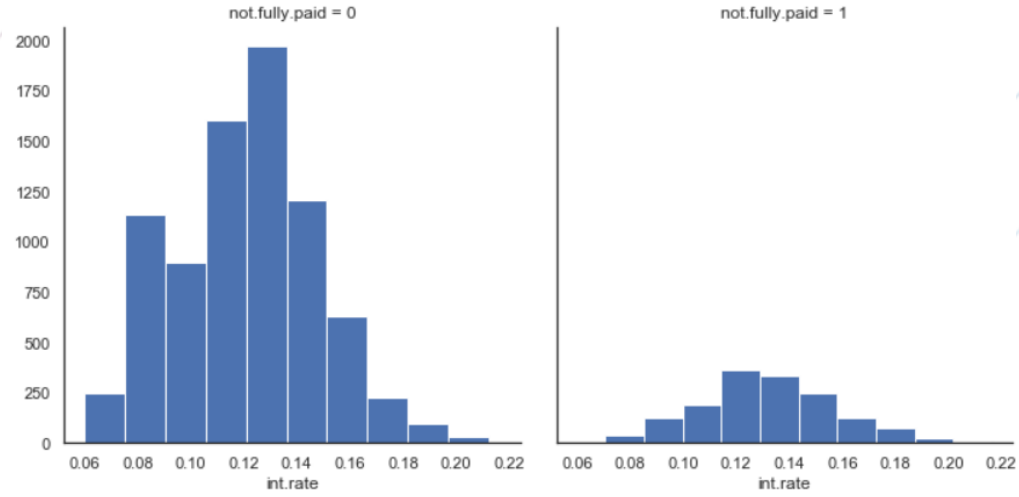
Separating the effect of not fully paid



Separating the effect of not fully paid

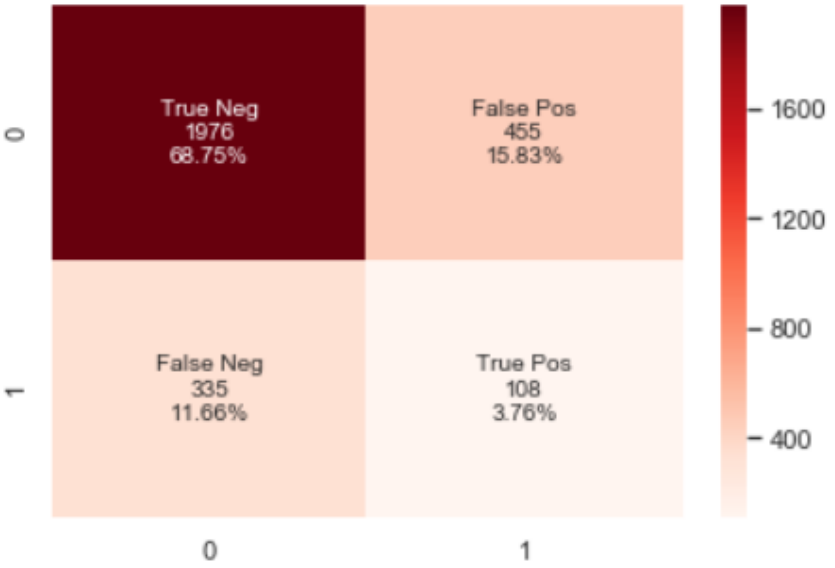




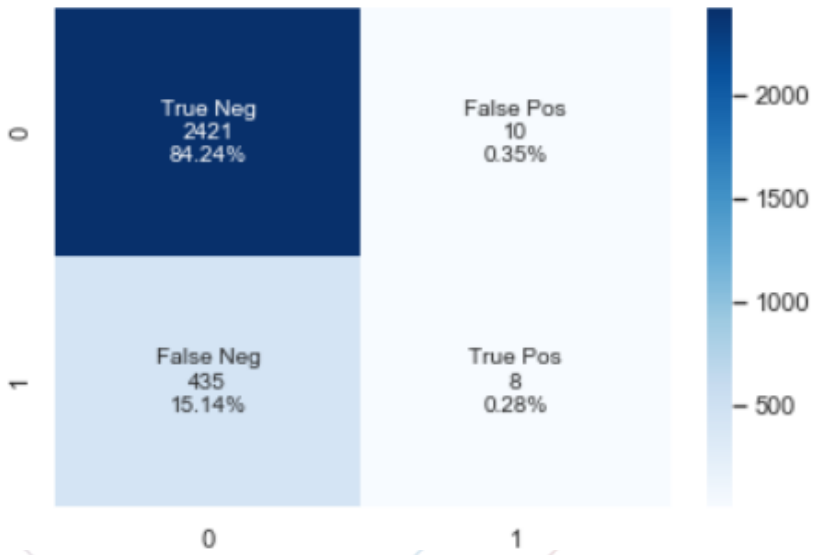


Which model do you think is better?

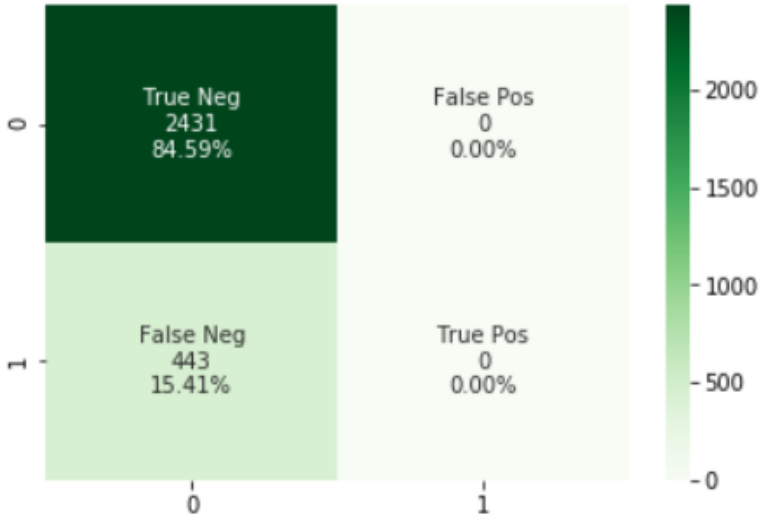
Decision tree



Random Forest

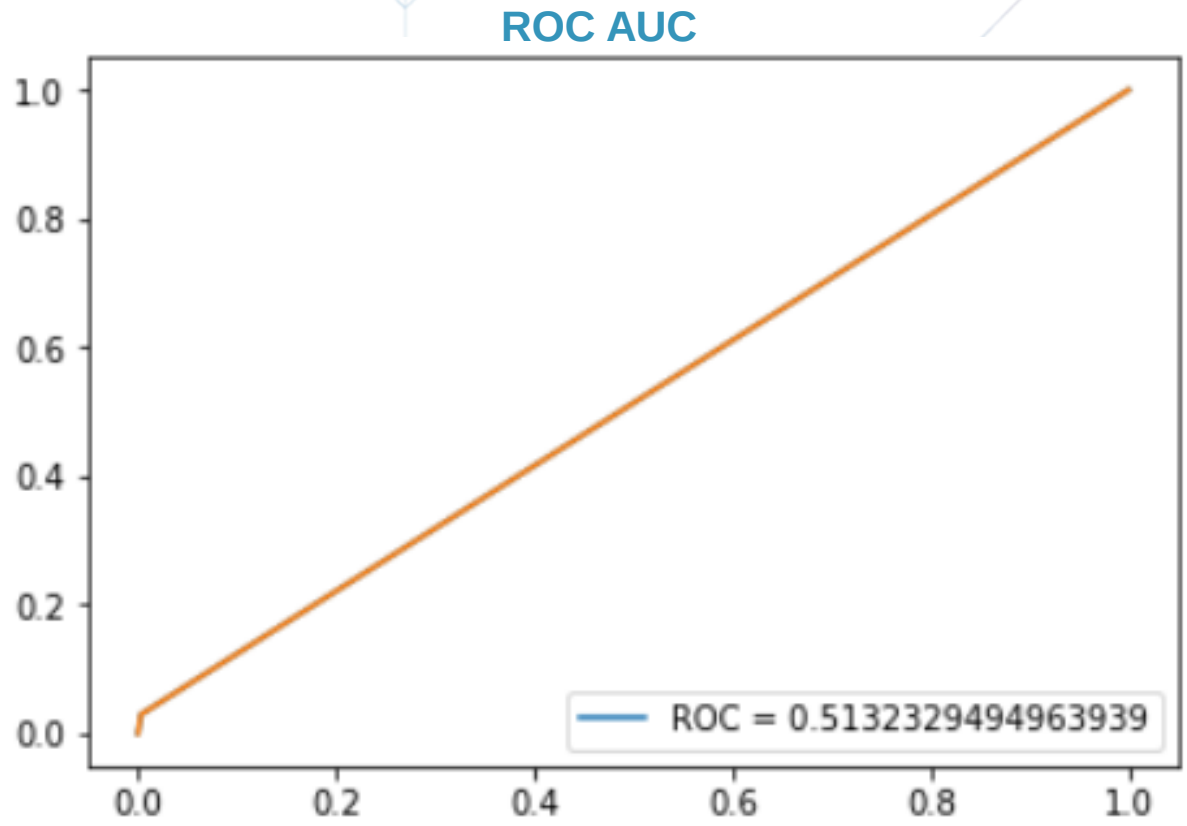


Grid Search



Mean Squared Error (MSE): 0.153
Root Mean Squared Error (RMSE): 0.39
R^2: 0.846

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 you!

Appendix