Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project>

# Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

## Key Decisions:

Answer these questions

* What decisions needs to be made?

**Soy funcionario de un banco, y soy responsable de determinar si los clientes son solventes para otorgarles prestamos, generalmente, el banco recibe 200 solicitudes de prestamos semanales y la aprobación es a mano.**

**Luego de un escandalo financiero que afectó al banco rival, de repente el equipo recepciona mas de 500 solicitudes de prestamos por semana. El gerente ve esto como una gran oportunidad y quiere que se descubra como procesar todas las solicitudes de préstamo en un plazo no mayor de una semana.**

**Según los modelos de clasificación que aprendí recientemente, necesito evaluar sistemáticamente las solvencia de estos nuevos solicitantes de préstamos, y proporcionarle al gerente la lista de clientes solventes.**

* What data is needed to inform those decisions?
* **El Dataset sobre todas las solicitudes de créditos pasados**
* **La lista de clientes que deben procesarse en los próximos días**
* What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

**Necesitamos utilizar el modelo binario para tomar nuestras decisiones, porque lo que buscamos es identificar a las personas que califican y no califican para prestamos bancarios.**

# Step 2: Building the Training Set

*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn’t* ***need to convert any data fields to the appropriate data types.***

*Here are some guidelines to help guide your data cleanup:*

* For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.
* Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
* Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.
* Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

***Note:*** *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

***Note:*** *For students using software other than Alteryx, please format each variable as:*

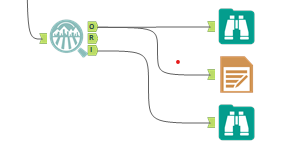
|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Credit-Application-Result | String |
| Account-Balance | String |
| Duration-of-Credit-Month | Double |
| Payment-Status-of-Previous-Credit | String |
| Purpose | String |
| Credit-Amount | Double |
| Value-Savings-Stocks | String |
| Length-of-current-employment | String |
| Instalment-per-cent | Double |
| Guarantors | String |
| Duration-in-Current-address | Double |
| Most-valuable-available-asset | Double |
| Age-years | Double |
| Concurrent-Credits | String |
| Type-of-apartment | Double |
| No-of-Credits-at-this-Bank | String |
| Occupation | Double |
| No-of-dependents | Double |
| Telephone | Double |
| Foreign-Worker | Double |

*To achieve consistent results reviewers expect.*

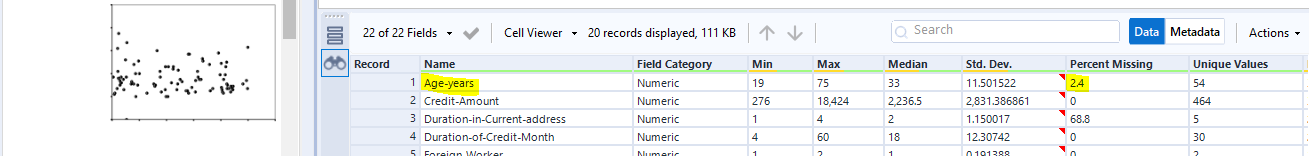
*Answer this question:*

* In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
* **Voy a imputar Age-Years**
* **Voy a eliminar Duration-in-Current-address, Occupation, Concurrent-Credits, Guarantors, Foreign-Worker, No-of-dependents, and Telephone.**

**Logré identifcar los datos faltantes y los campos de baja variabilidad usando la herramienta “Field\_Summary” y analizando el informe de salida interactiva.**

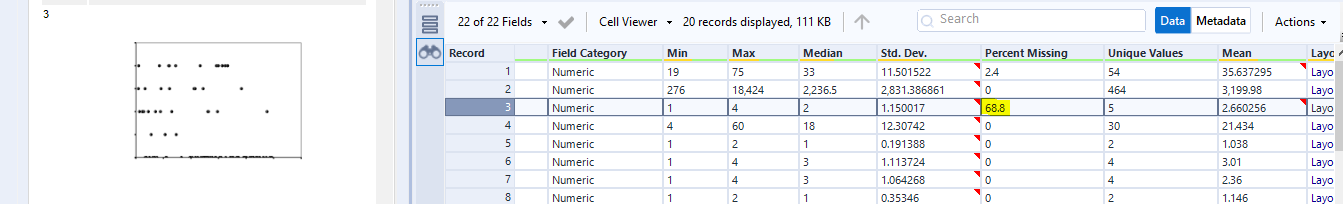


**Datos Perdidos**



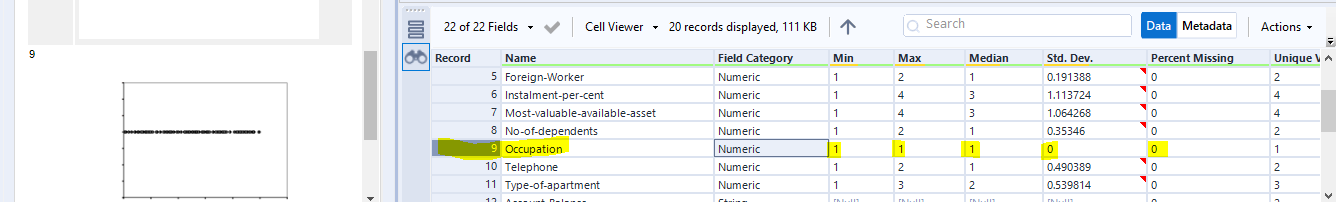
**El campo Age-Years tiene 2.4 porciento de datos faltantes, al usar la herramienta “Field\_Summary” obtuve el resultado, voy a imputar el campo utilizando la mediana de todo el campo.**

**El campo Duration-in-Current-address tiene un 68.8% de datos faltantes, por ende serán eliminados**

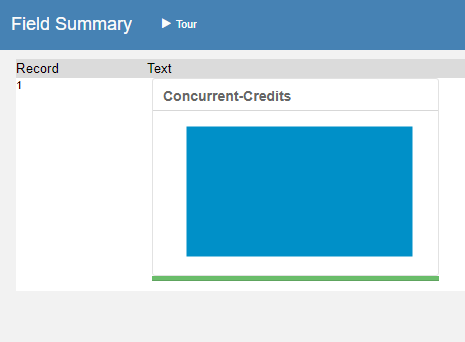


# **Existe baja variabilidad: nuestros datos son uniformes y no existe variacion de los datos.**

**El campo ocupación solo tiene “1” como valor**

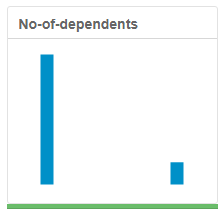


**El campo creditos concurrentes solo tiene un valor “Otros bancos / depósitos”, 500 instancias en total.**

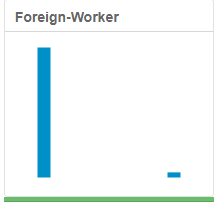


# Baja variabilidad: el campo de datos tiene sesgo hacia un tipo de datos

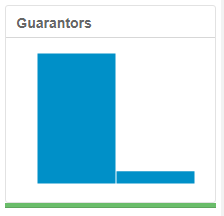
* + - * El campo de teléfono debe eliminarse porque no aporta nada sobre la solvencia.
      * No de dependientes debe eleminarse porque hay sesgo hacia el “1”



* + - * Trabajador extranjero debe eliminarse porque la mayoría de los datos está sesgado en “1”.



* + - * Garantes la mayoría de los datos están sesgados hacia el “1”



# Step 3: Train your Classification Models

*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.*

*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

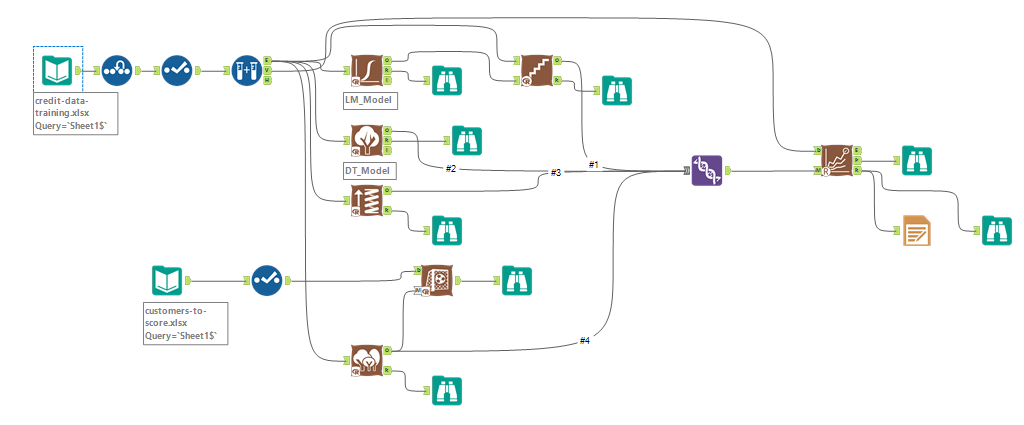
*Answer these questions for* ***each model*** *you created:*

* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

*You should have four sets of questions answered. (500 word limit)*

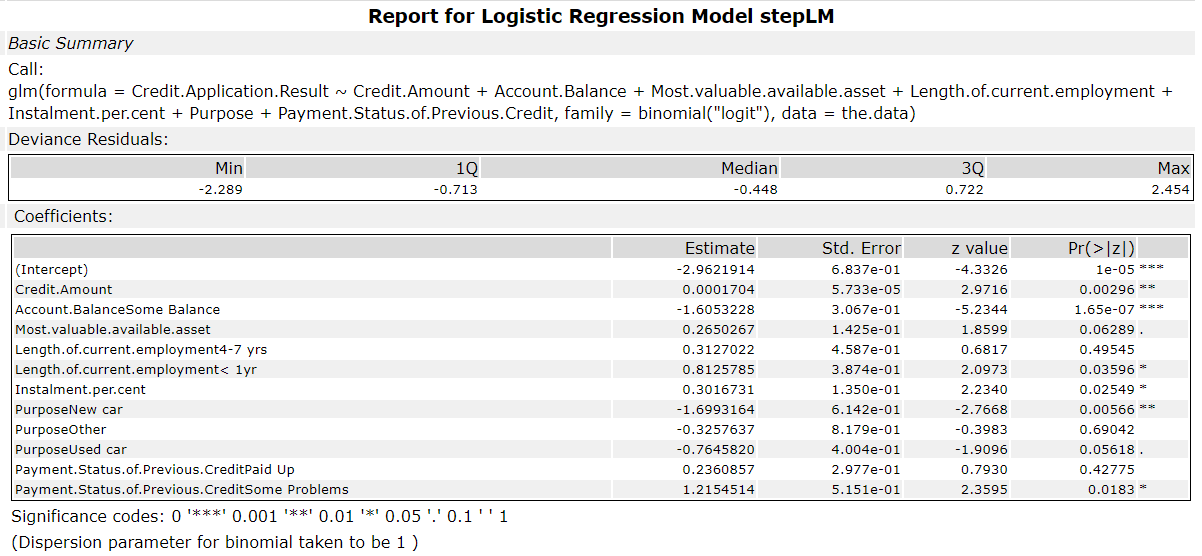
***Hice los cuatro modelos en Alteryx y usé la herramienta para comparar los modelos para validar***

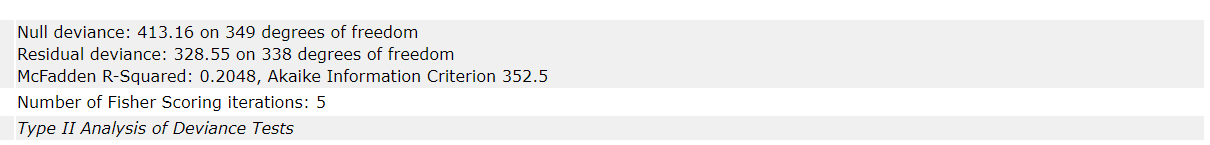
***Logistic Stepwise, Decision Tree, Forest Model y Boosted Model.***

******

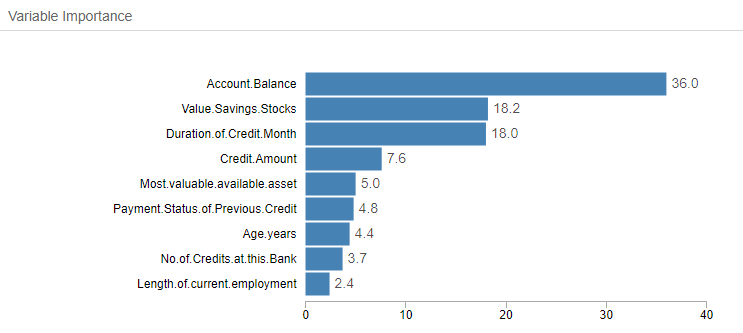
***Variables predictivas significativas***

***Logistic\_Stepwise: segun el reporte, las variables predictivas significativa son saldo de la cuenta, el estado de pago del crédito anterior, el propósito, el monto del crédito, la duración del trabajo actual y porcentaje de pagos a plazos.***





*Arbol de decisión: según el reporte de importancia de variables que se muestra en el grafico, las 3 principales variables predictivas son el saldo de la cuenta, las acciones de ahorro de valor y la duración del mes de crédito.*



***Modelo Forestal: se observa en el grafico de importancia, las tres principales variables predictivas son Monto del crédito, años de antigüedad y duración del mes de crédito***

# Step 4: Writeup

*Decide on the best model and score your new customers. For reviewing consistency, if Score\_Creditworthy is greater than Score\_NonCreditworthy, the person should be labeled as “Creditworthy”*

*Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)*

*Answer these questions:*

* Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
  + Overall Accuracy against your Validation set
  + Accuracies within “Creditworthy” and “Non-Creditworthy” segments
  + ROC graph
  + Bias in the Confusion Matrices

**Note:** Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

* How many individuals are creditworthy?

**Before you Submit**

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/265/view) here. Reviewers will use this rubric to grade your project.