Stephane Keil Rios
160559

Aprendizaje de Maquina

Proyecto Final: Clasificación del ingreso en función a variables censales

1. Antecedentes:

La economia del siglo XXI en los países de alto ingreso como en los paises de las economias en transición se sustenta en gran parte del consumo de los hogares. Los nuevos canales de distribución y las tecnologias de las información han transformado radicalmente la manera en la que consumimos, al menos para los niveles socioeconomicos más altos. El sinumero de productos y servicios que se ofrecen a través de canales más diversos y más cercanos al consumidor final van dirigidos a un segmento meta o “target” cada vez más preciso. Las grandes empresas de investigación de mercado buscan proveer a las empresas de un conocimiento cada vez más profundo sobre sus consumidores. Una de las aplicaciones de este tipo de segmentaciones es determinar si una persona tiene la capacidad de compra para adquirir un producto o servicio antes de realizar un esfuerzo de venta que podria ser un gasto innecesario para las empresas. En este sentido el marketing dirigido o “Below the Line” requiere de determinar la capacidad de compra de un consumidor con información indirecta respecto a su ingreso y a sus gastos. Aunque hoy en día se cuenta cada vez más con información de multiples fuentes para detectar prospectos, las tecnicas principales de clasificación del Nivel Socioeconomico siguen basandose en ciertas caracteristicas del hogar y de la persoona que por lo general se obtienen de censos o encuestas. Dado que existe una importante preocupación respecto a la privacidad de los datos de las personas, es importante mantener la capacidad de estimar el nivel de ingreso de las personas con información que se puede conseguir sin necesidad de invadir la privacidad de las personas. En este sentido, es interesante revisar como las tecnicas estadisticas modernas y de aprendizaje de maquina pueden mejorar la capacidad de clasificar a un potencial comprador a su nivel socioeconomico más probable con la menor cantidad de información posible. Aunque no es un problema nuevo ni mucho menos nos proponemos analizar los resultados que diferentes tecnicas de aprendizaje de maquina pueden

proveer a este problema revisando un problema bien conocido y estudiado en la literatura.

2. Metas del problema asociado a resolver:

El objetivo de este analisis es aplicar tecnicas de mineria de datos y de aprendizaje de maquina sobre datos demograficos para identificar individuos con alta probabilidad de ganar más de 50,000 Dolares Estadounidienses (USD) por año en base a caracteristicas de los individuos como su edad, genero, nivel educativo, ocupación, etnicidad y país de origen. Buscando consolidar el aprendizaje de nuevas tecnicas de analisis de datos, decidimos revisar un problema bien estudiado proveniente del UCI Machine Learning Dataset Repository y corresponde a información extraida del censo de 1994. Nos proponemos realizar un analisis supervisado sobre esta información para la generación del mejor modelo de clasificación para saber si un individuo gana más o menos de 50,000 USD anuales.

3. Conjunto de datos:

El conjunto de datos contiene información anonimizada de 48,842 individuos para 14 variables de interes (predictores), además de tener una clasificación sobre si el individuo gana más de 50,000 USD por año (variable objetivo).

Atributo	**Descripción**	**Tipo**
age	Edad	Cuantitativo - Discreto
workclass	Tipo de empleador (Sector privado, gobierno, autoempleado, desempleado)	Categorico - Nominal
fnlwgt	Factor de expansión censal	Cuantitativo - Continuo
education	Maximo grado de estudios alcanzado - nombre del ultimo grado terminado	Categorico - Nominal
education_num	Maximo grado de estudios alcanzados en años de estudio	Cuantitativo - Discreto
marital_statu	Estado civil	Categorico -

s		Nominal
occupation	Tipo de empleo (Administrativo, Agricultura, Servicios, Industria)	Categorico - Nominal
relationship	Tipo de relación (Esposo, Soltero)	Cateogrico - Nominal
race	Etnicidad	Categorico - Nominal
sex	Sexo biologico	Categorico - Nominal
capital_gain	Ganacia de capital	Cuantitativo - Continuo
capital_loss	Perdida de capital	Cuantitativo - Continuo
hours_per_week	Numero de horas trabajadas a la semana	Cuantitativo - Discreto
native_country	País de origen	Categorico - Nominal
earning_class	Clase de ingresos	Categorico - Nominal

4. Analisis Exploratorio de los Datos:

A continuación se muestran los histogramas de todas las variables en el conjunto, es interesante notar que existe mucha homogeneidad en los datos, 85% de los individuos son de etnicidad blanca, 90% de los individuos nacieron en los Estados Unidos, el 68% trabaja en el sector privado, el 66% son hombres y casi la totalidad de los individuos no mostraron ganancias o perdidas de capital.

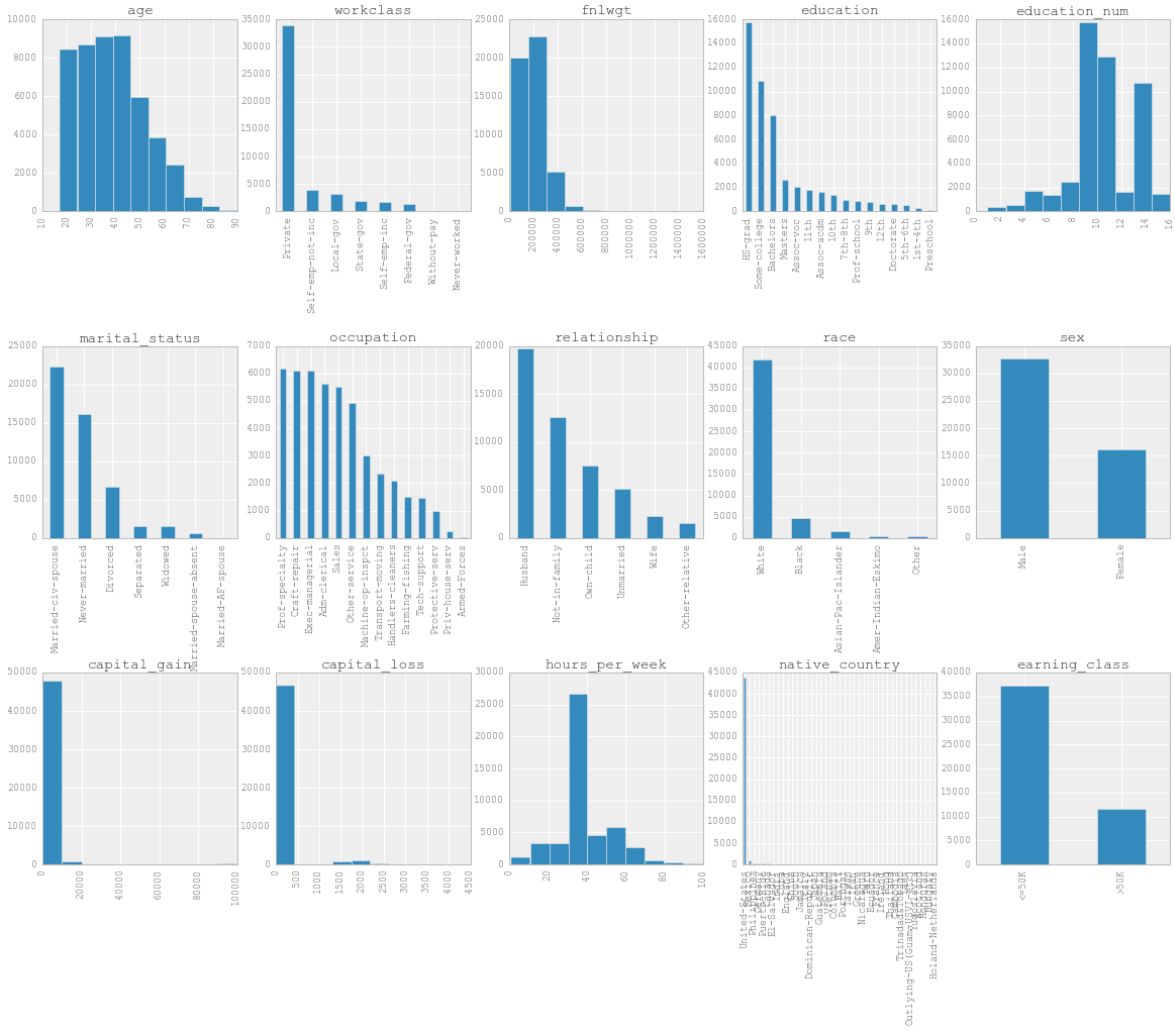


Figura 1 – Histogramas para conteo de frecuencias de los 15 atributos del conjunto de datos.

5. Preprocesamiento y transformación de los datos

Dado que la variable de **fnlwgt** es un factor de expansión podemos considerar que no sera de utilidad en el analisis ya que no describe ninguna caracteristica de la población, de la misma forma las variables education y education_num son en realidad la misma variable codificada de forma diferente. Se decidió eliminar la variable **education_num** ya que no existe ningun motivo para pensar que el efecto entre cada nivel de estudios completado es el mismo entre todos los niveles de estudio por lo que se mantuvo la variable education como categorica.

a. Valores faltantes

Se revisó la información para determinar si existián valores faltantes, se encontró que tres de las variables categoricas contienen elementos faltantes.

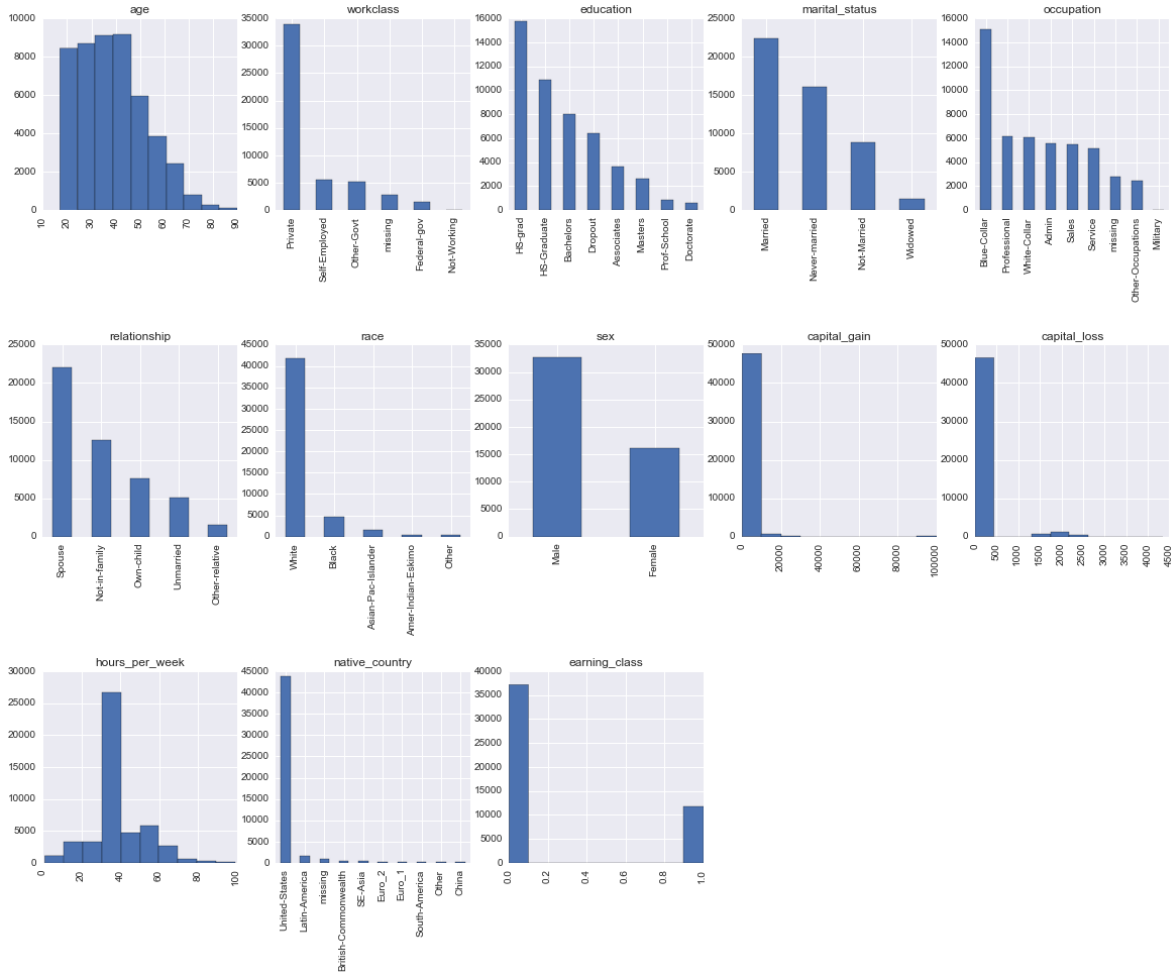
Variable	Valores Faltantes	Proporción
workclass	2,799	5.7%
occupation	2,809	5.8%
native_country	857	1.8%

Dado que los valores faltantes unicamente se presentan en el caso de variables categoricas y a falta de un analisis más detallado se tomo la decisión de clasificar los valores faltantes en una categoria adicional que llamaremos ‘missing’ y dejaremos dentro del analisis.

b. Agrupación de Categorias

Dado que se tiene un gran numero de categorias parecidas entre ellas se tomo la decisión de recodificar aquellas que son muy similares, este ejercicio permitio simplificar considerablemente la matriz de diseño. Se pueden revisar los cambios de variables realizados en el Apendice 1.

Después de las tranformaciones realizadas, los datos quedan como sigue:



c. Codificación de las variables categoricas

La codificación de las variables categoricas se realizo con variables pseudobinarias dónde por cada valor que puede tomar una categoria se genera una nueva columna con un valor de 0 si la observación no toma el valor y 1 si la observación toma el valor.

d. Estrategia de Validación Cruzada

Se decidio estudiar la bondad de ajuste de los modelos con el uso de una estrategia de validación cruzada simple dejando 80% de los datos en el conjunto de entrenamiento y 20% en el conjunto de validación.

e. Estandarización de los datos

Varios de los metodos que probamos requieren que la información este en una escala similar, de no ser así podria darse el caso de dar mayor importancia a ciertas variables por la simple razon de la escala en la que se encuentra clasificada. Utilizamos el escalador estandar de Python que normaliza la información con media 0 y desviación estandar de 1.

6. Analisis de los datos:

Dado que estamos trabajando con datos socio-demograficos para tratar de definir el Nivel Socioeconomico de un individuo es importante entender de entrada cuales son las variables que tienen un efecto sobre nuestra variable objetivo. Un ejercicio muy sencillo es obtener el indice de correlación de Pearson para ver cuales de nuestras variables muestran un efecto predictivo lineal sobre nuestra variables objetivo. A continuación se muestra la correlación de todos nuestros atributos con la variable de clase de ingreso.

Variable	Corr	Variable	Corr
relationship_Spouse	0.45	native_country_missing	0.01
marital_status_Married	0.45	occupation_Military	0.00
age	0.23	native_country_SE-Asia	0.00
hours_per_week	0.23	native_country_Euro_2	0.00
capital_gain	0.22	workclass_Not-Working	-0.01
sex_Male	0.21	race_Other	-0.02
occupation_White-Collar	0.21	race_Amer-Indian-Eskimo	-0.03
occupation_Professional	0.19	native_country_South-America	-0.03
education_Bachelors	0.18	education_HS-Graduate	-0.06
education_Masters	0.17	marital_status_Widowed	-0.07
education_Prof-School	0.15	native_country_Latin-America	-0.08
capital_loss	0.15	workclass_Private	-0.08
education_Doctorate	0.13	workclass_missing	-0.08
workclass_Self-Employed	0.10	occupation_missing	-0.08
race_White	0.08	relationship_Other-relative	-0.09
workclass_Federal-gov	0.06	occupation_Admin	-0.09
workclass_Other-Govt	0.04	race_Black	-0.09
native_country_United-States	0.03	occupation_Blue-Collar	-0.11
occupation_Other-Occupations	0.03	education_HS-grad	-0.13
native_country_British-Commonwealth	0.03	relationship_Unmarried	-0.14
occupation_Sales	0.02	marital_status_Not-Married	-0.16
native_country_Euro_1	0.01	occupation_Service	-0.16

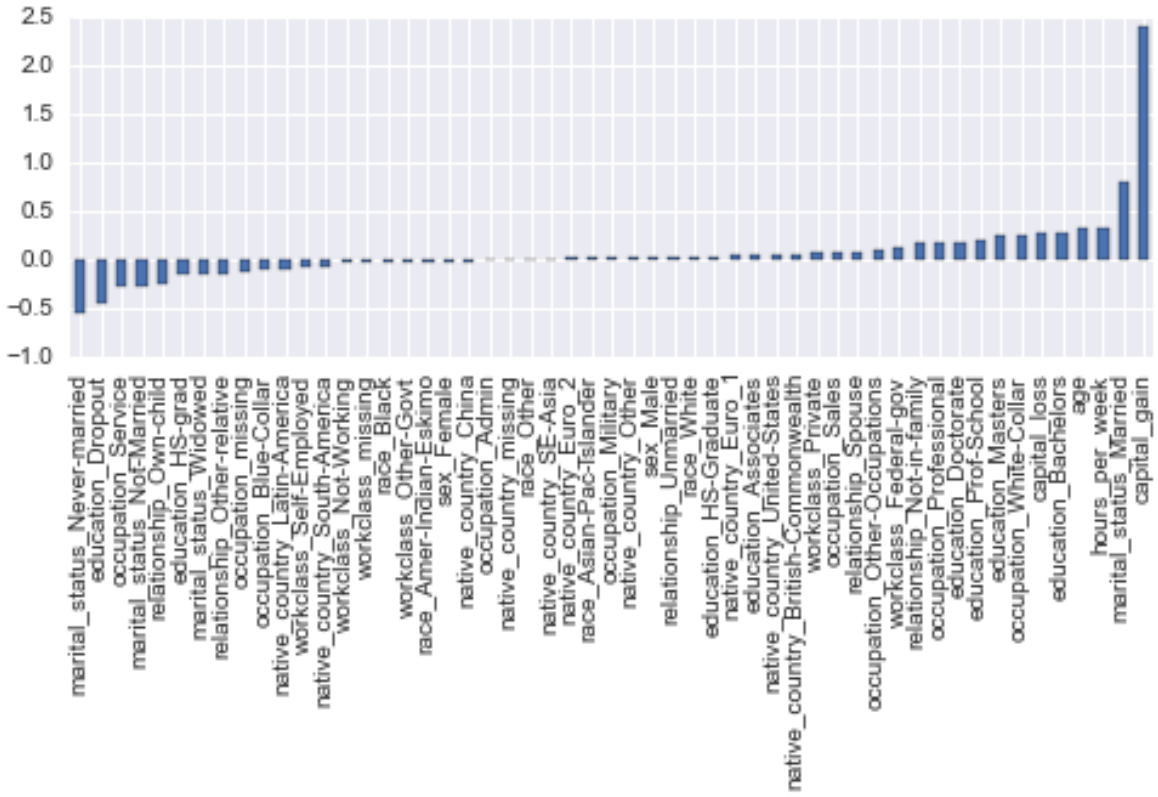
native_country_China	0.01	education_Dropout	-0.17
race_Asian-Pac-Islander	0.01	relationship_Not-in-family	-0.19
native_country_Other	0.01	sex_Female	-0.21
education_Associates	0.01	relationship_Own-child	-0.23
		marital_status_Never-married	-0.32

Podemos observar que los individuos casados, de mayor edad, con ocupaciones de profesionista, que trabajan un mayor numero de horas, de sexo masculino y con mayores niveles educativos tienden a mostrar mayor posibilidad de pertenecer a la clase que gana más de 50 K USD. Las variables de Ganancia de capital y de perdida de capital también muestran una relación positiva con pertenecer a la clase más afluente. La intuición nos indica que aquellos individuos que tienen inversiones, aunque estas les hallan generado perdidas, tienden a tener mayor nivel de ingresos que los que no. Al contrario, los individuos solteros, de sexo femenino, de etnicidad afroamericana, con menores niveles educativos y con ocupaciones de tipo industrial o obrera tienden a estar más asociados con la clase de menores ingresos.

a. Regresión logistica

Una simple regresión logistica puede darnos información sobre la complejidad de nuestro problema, este metodo tiene la gran ventaja de que nos permite revisar cuales son los predictores que más influyen sobre la variable objetivo. Obtuvimos un de **84.6%** sobre el conjunto de prueba y la matriz de confusión normalizada puede verse como sigue:

Con este clasificador unicamente estamos logrando clasificar el 61% de los individuos que ganan más de 50 K en la clase adecuada. Veamos ahora cuales son los coeficientes que más pesan en este clasificador.



De la grafica cabe resaltar que ganancia de capital, estado civil – casado, horas laboradas a la semana, edad, nivel educativo superior y una ocupación de tipo profesional incrmentan la probabilidad de ganar más de 50 K por año mientras que las variables de estado civil – soltero, no haber terminado la preparatoria, trabajar en ocupación de servicio y tener una ocupación industrial o de tipo obrero reducen la probabilidad de pertenecer a la clase de mayor ingreso.

b. Aplicación de los metodos:

Debido a la naturaleza de la información analizada decidimos utilizar metodos que funcionan bien sobre variables categoricas como lo son los arboles de decisión y el metodo ingenuo de Bayes, incluimos un metodo de ensamble llamado bosques aleatorios. Aunque no es un metodo que funcione bien en altas dimensiones decidimos probar con el metodo de k-vecinos más cercanos. Queriamos probar con metodos que permiten encontrar buenas aproximaciones para relaciones no lineales por lo que incluimos las maquinas de soporte vectorial y aunque quisimos probar con el uso de redes neuronales, los largos tiempos de entrenamiento no nos permitieron realizar pruebas satisfactorias. Se utilizaron para esta fase niveles de complejidad medios que se indican en cada uno de los metodos. Cabe mencionar que todos los metodos dan resultados similares con un accuracy_score alrededor del **85%.**

Metodo	Complejidad	Accuracy	F1 Score
Maquinas de Soporte Vectorial	kernel=RBF C=6.0	84.69%	0.6416
Bosques aleatorios	Arboles=20	84.19%	0.6414
Decision Tree	Max Depth=20	84.65%	0.6601
Naive Bayes	Gaussiana	76.59%	0.6258
Regresión Logistica		84.59%	0.6446
K vecinos mas cercanos	K=10	83.42%	0.5927

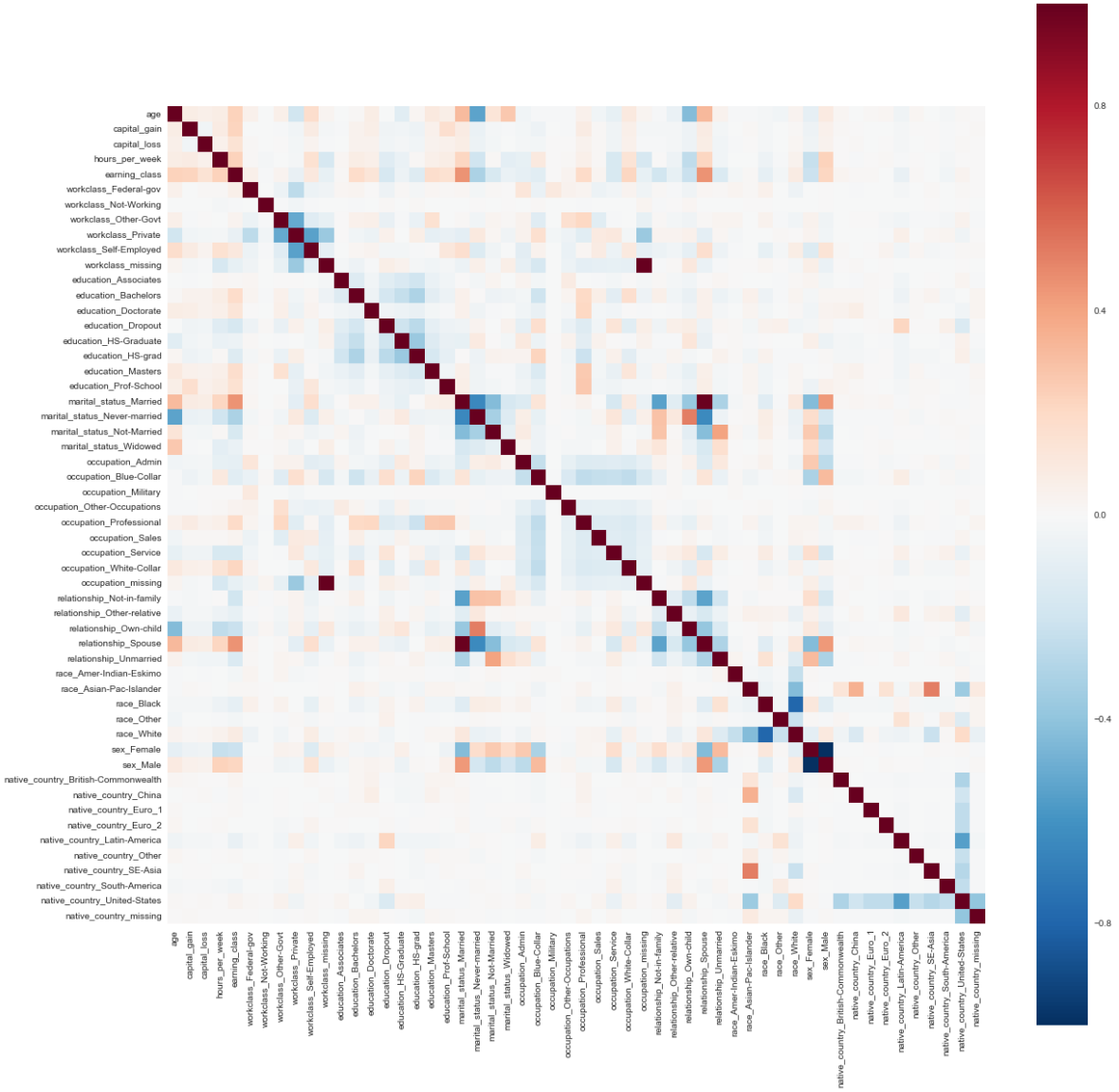
Es evidente de estos resultados que sin mayor transformación de los datos de entrada, no vamos a poder mejorar la precisión de nuestro modelo.

7. Selección de atributos

Para mejorar la precisión de los metodos aplicados es indispensable reducir el número de dimensiones con el que se trabaja. Intentamos mantener la interpretabilidad del modelo aplicando un simple analisis de correlaciones entre las variables y dejamos como segunda estrategia el uso de un metodo más elaborado de Analisis de Componentes Principales. La tercera estrategia fue un algoritmo de Greedy Forward Search modificado.

a. Analisis de correlación

Hemos observado cierta redundancia en nuestras variables predictoras, es momento de simplificar nuestra matriz de diseño y realizar un primer ejercicio de reducción de dimensionalidad. Al observar el siguiente grafico de correlaciones, podemos observar que existen ciertas variables redundantes que pueden ser eliminadas dado que explican lo mismo. Algunas de estas relaciones son evidentes, como por ejemplo la correlación entre las variables sex_male y sex_female pero existen otras menos evidentes como marital_status_married y relationship_spouse.



Vamos a identificar todas las variables con más de 50% de correlación entre ellas y quedarnos solo con aquellas que nos parece tener mayor poder de comunicación de las relaciones que ocurren. Al realizar esta operación eliminamos las siguientes variables:

"sex_Female","workclass_missing","relationship_Spouse","race_White","marital_status_Never-married","native_country_United-States","relationship_Not-in-family","workclass_Private","race_Asian-Pac-Islander"

Al volver a correr los modelos se logro un ligero incremento en la precisión (accuracy) de los modelos, sin embargo estos son marginales (Accuracy de Logistic Regression subio a tan solo **85.3%**). Se intentaron diferentes valores de unmbral para la selección de variables, sin embargo esto no represento mejoras importantes y tuvimos que descartar esta estrategia.

b. PCA

Decidimos realizar pruebas diferentes, aplicar el metodo de descomposición en componentes principales que conserven cierta cantidad de varianza en los datos, probamos este ejercicio con la regresión logistica debido a la velocidad con la que entrega resultados. Se generaron modelos de regresión logistica para valores de varianza explicada por las componentes principales desde 5% hasta 95% en incrementos de 5%. Los resultados en cuanto a la precisión del modelo se presentan a continuación:

Componentes principales	Varianza explicada	Accuracy	F1 Score
1	6.3%	80.7%	0.513311
2	11.2%	82.0%	0.552899
3	15.4%	82.0%	0.553462
5	22.1%	82.4%	0.575099
6	25.1%	82.7%	0.584934
8	31.0%	82.7%	0.586621
10	36.4%	82.8%	0.587255
12	41.5%	82.7%	0.588521
14	46.4%	82.8%	0.589325
16	51.2%	82.7%	0.590005
18	55.9%	82.9%	0.593439
20	60.5%	83.0%	0.596065
22	65.1%	83.0%	0.596942
25	71.8%	83.0%	0.597232
27	76.2%	83.0%	0.598058
29	80.5%	83.2%	0.601895
32	86.5%	83.6%	0.609946
34	90.1%	84.6%	0.628965
38	96.3%	84.9%	0.637774

De la tabla podemos observar que el metodo de descomposición no muestra mejoras sustanciales al modelo original con todas las variables.

c. Greedy Forward Search

Realizamos una tercera prueba utilizando un algoritmo de greedy forward search modificado. Para ello tomamos el modelo de regresión logistica que se

presentó en la sección 6.a del presente documento. Al ordenar los valores absolutos de los coeficientes obtenidos de la regresión, decidimos probar el modelo de regresión logistica añadiendo uno a uno las variables cuyo coeficiente muestra mayor poder explicativo respecto a nuestra variable objetivo.

Atributos	Accuracy	F1 Score
1	0.795987	0.320027
2	0.791381	0.316107
3	0.792916	0.314934
4	0.795475	0.328629
5	0.796192	0.433248
6	0.802232	0.49635
7	0.821169	0.565315
8	0.831815	0.598975
9	0.832941	0.602339
10	0.838366	0.621977
11	0.839902	0.623858
12	0.844508	0.63721
13	0.848807	0.649917
14	0.8485	0.649953
15	0.849012	0.651547
16	0.849524	0.656702
17	0.847477	0.652194
18	0.847886	0.652804
19	0.848193	0.65407
20	0.847784	0.652651
21	0.849729	0.658446
22	0.84891	0.656904
23	0.849012	0.657056
24	0.849626	0.658927
25	0.849831	0.659233
26	0.849933	0.659703
27	0.849012	0.658328
28	0.849729	0.660656
29	0.850445	0.66204
30	0.850957	0.663586
31	0.850957	0.663586
32	0.850036	0.661115
33	0.850138	0.661425
34	0.850036	0.661272
35	0.850138	0.661581

36	0.850241	0.66189
37	0.850445	0.66204
38	0.850752	0.662812
39	0.850752	0.662812
40	0.850752	0.662812
41	0.850752	0.662812
42	0.850752	0.662812
43	0.85065	0.662659

Aunque las mejoras en terminos de precisión continuan siendo marginales, podemos observar que el valor maximo se alcanza cuando utilizamos los 30 atributos con mayor poder explicativo de la regresión logistica. Dado que de las tres estrategias utilizadas esta fue la que mejores resultados nos brindo, vamos a utilizarla con los demas modelos para la selección del mejor clasificador para nuestro ejercicio.

8. Selección de un clasificador

a. Selección de clasificador y complejidad del modelo

Para poder seleccionar el mejor clasificador es necesario explorar la complejidad de cada uno de los algoritmos seleccionados. Cada algoritmo expresa su complejidad de manera diferente y no son comparables entre ellas por lo que se decidio probar cada uno de los algoritmos con multiples niveles de complejidad. Esto permite que podamos comparar cada uno de los modelos entre ellos observando unicamente las metricas de precisión que hemos estado utilizando. A continuación se presentan los resultados para cada uno de los modelos utilizados:

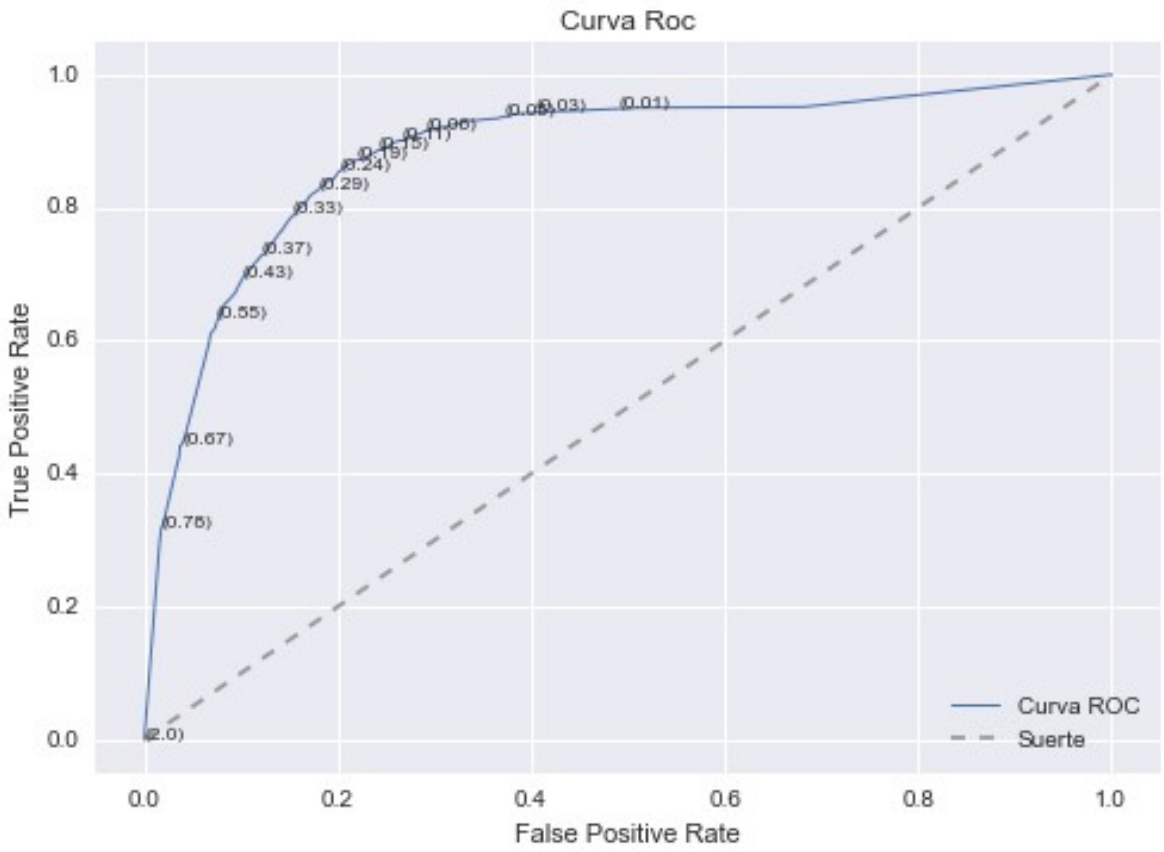
Algoritmo	Complejidad	Accuracy	F1 Score
Bayes Ingenuo	Gaussiano	78.0%	0.641597
Regresión Logistica	C=1	84.9%	0.66036
K Vecinos	k=1	80.9%	0.605006
	k=6	83.2%	0.597205
	k=11	83.6%	0.632518
	k=16	83.7%	0.622726
	k=21	83.6%	0.638952
Bosques	arboles= 10	84.4%	0.653794

Aleatorios	arboles= 20	84.5%	0.655634
	arboles= 30	84.5%	0.655938
	arboles= 40	84.4%	0.65673
	arboles= 50	84.6%	0.659415
	arboles= 60	84.6%	0.659892
	arboles= 70	84.5%	0.657162
	arboles= 80	84.6%	0.658365
	arboles= 90	84.6%	0.65852
Arbol de Decisión para Clasificación	criterio=entropia profundidad=5	81.5%	0.652716
	criterio=entropia profundidad=10	85.3%	0.677332
	criterio=entropia profundidad=15	85.0%	0.67352
	criterio=entropia profundidad=20	84.4%	0.665791
	criterio=entropia profundidad=25	83.6%	0.659284
	criterio=gini profundidad=5	81.5%	0.654008
	criterio=gini profundidad=10	85.3%	0.676018
	criterio=gini profundidad=15	85.2%	0.682627
	criterio=gini profundidad=20	84.4%	0.678459
	criterio=gini profundidad=25	83.4%	0.653731
Maquinas de Soporte Vectorial	kernel= radial gamma=1	82.8%	0.578934
	kernel= radial gamma=6	81.0%	0.488263
	kernel= radial gamma=11	80.3%	0.441287
	kernel= radial gamma=16	79.7%	0.408236

De todos los modelos probados los que tuvieron mejor desempeño fueron la regresión logistica, los bosques aleatorios y los arboles de decisión. El mejor modelo en terminos del F1 Score fue el de Arbol Binario de Clasificación con criterio de Gini y profundidad de 15 niveles con Accuracy de 85.2% y F1 Score de 0.68. Es interesante notar que al final todos los algoritmos generan resultados bastante similares (siempre y cuando los niveles de complejidad) sean comparables. Utilizaremos este algoritmo para generar nuestro clasificador. Una de sus caracteristicas más interesantes es que es facil producir las reglas para incorporarlo a los sistemas de producción de una empresa, al final es solo una secuancia larga de estructuras condicionales de control.

b. Selección del umbral de clasificación

Una vez seleccionado el modelo es importante determinar el umbral de clasificación que mejor se ajusta a los propositos del ejercicio. Para nuestro problema vamos a asumir que una empresa quiere promocionar un producto de lujo y sabe que su target esta relacionado con personas que ganan más de 50,000 USD por año. Aunque quieren minimizar el costo de contactar a personas que no cumplen con el nivel socioeconomico de su target, quieren asegurar que se contacta a la mayor cantidad de individuos potenciales, por lo que la proporción de falsos positivos tiene menor importancia que los falsos negativos, uno de los ejecutivos de la empresa, que estudio un MBA en una Universidad prestigiosa constantemente usa el mantra del 80/20, por lo que se desea contactar al menos al 80% de los potenciales compradores del nuevo producto. Utilizaremos una curva ROC para determinar el umbral que asegura contactar al menos al 80% de los individuos de la clase con nivel adquisitivo mayor. Por lo que se toma el valor de umbral que cumple con una tasa de verdaderos positivos del 80%. A continuación se presenta la curva ROC con los valores anotados:



Podemos observar que el 80% de los verdaderos positivos se alcanza con un umbral de clasificación de 0.33 es decir que cuando la probabilidad de que un individuo sea de la clase que gana más de 50K USD por año podemos asumir que pertenece a esta clase y enviarle la publicidad para el nuevo producto. Revisemos cuales son las implicaciones de esta decisión.

9. Conclusiones

Nuestro clasificador utilizando el algoritmo de Arbol de Decisión de Clasificación con un umbral de 33% nos indica que tenemos una precisión en la clasificación del 83.2% con un F1 Score de 0.6948.

	precision	recall	f1-score	casos
<50K	0.93	0.85	0.88	7396
>50K	0.62	0.79	0.69	2373
total	0.85	0.83	0.84	9769

Logramos identificar 79% de los individuos de la clase alta (hay 20% de potenciales compradores que no seran contactados). De los individuos contactados unicamente el 62% de estos realmente pertenece a la clase que nos interesa es decir que hay un 38% de los individuos contactados que no corresponden a la clase alta.

10. Referencias

- UCI Archive, 2011, http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names
- Mihov, Valentin. (2015, April 17). Adult Income Data Set Analysis with Ipython, Random Pieces of Wisdom about Technology. [Blog Entry]. Retrieved from https://www.valentinmihov.com/2015/04/17/adult-income-data-set/
- Hofmann, Markus. (2011, May 21). Predicting earning potential on Adult Dataset. Blanchardstown Institute of Technology
- Statistical Consulting Group. (2013, September 25). Data Set: Adult (R). San Diego University. [Blog Entry] Retrieved from: http://scg.sdsu.edu/dataset-adult_r/
- Layton, Robert. (2015). Learning Data Mining with Python. Birmhingam, UK: Packt Publishing.
- Kuhn, Max & Johnson, Kjell. (2013). Applied Predictive Modeling. New York, USA: Springer.

11. Apendice 1 – Cambio de Variables

	Valor Original	Transformación
Workclass	State-gov	Other-Govt
Workclass	Local-gov	Other-Govt
Workclass	Self-emp-inc	Self-Employed
Workclass	Self-emp-not-inc	Self-Employed
Workclass	Without-pay	Not-Working
Workclass	Never-worked	Not-Working
Occupation	Adm-clerical	Admin
Occupation	Armed-Forces	Military
Occupation	Craft-repair	Blue-Collar
Occupation	Exec-managerial	White-Collar
Occupation	Farming-fishing	Blue-Collar
Occupation	Handlers-cleaners	Blue-Collar
Occupation	Machine-op-inspct	Blue-Collar
Occupation	Other-service	Service
Occupation	Priv-house-serv	Service
Occupation	Prof-specialty	Professional
Occupation	Protective-serv	Other-Occupations
Occupation	Sales	Sales
Occupation	Tech-support	Other-Occupations
Occupation	Transport-moving	Blue-Collar
Native Country	Cambodia	SE-Asia
Native Country	Canada	British-Commonwealth
Native Country	China	China
Native Country	Columbia	South-America
Native Country	Cuba	Other
Native Country	Dominican-Republic	Latin-America
Native Country	Ecuador	South-America
Native Country	El-Salvador	South-America
Native Country	England	British-Commonwealth
Native Country	France	Euro_1
Native Country	Germany	Euro_1

Native Country	Greece	Euro_2
Native Country	Guatemala	Latin-America
Native Country	Haiti	Latin-America
Native Country	Holand-Netherlands	Euro_1
Native Country	Honduras	Latin-America
Native Country	Hong	China
Native Country	Hungary	Euro_2
Native Country	India	British-Commonwealth
Native Country	Iran	Other
Native Country	Ireland	British-Commonwealth
Native Country	Italy	Euro_1
Native Country	Jamaica	Latin-America
Native Country	Japan	Other
Native Country	Laos	SE-Asia
Native Country	Mexico	Latin-America
Native Country	Nicaragua	Latin-America
Native Country	Outlying-US(Guam-USVI-etc)	Latin-America
Native Country	Peru	South-America
Native Country	Philippines	SE-Asia
Native Country	Poland	Euro_2
Native Country	Portugal	Euro_2
Native Country	Puerto-Rico	Latin-America
Native Country	Scotland	British-Commonwealth
Native Country	South	Euro_2
Native	Taiwan	China

Country		
Native Country	Thailand	SE-Asia
Native Country	Trinadad&Tobago	Latin-America
Native Country	United-States	United-States
Native Country	Vietnam	SE-Asia
Native Country	Yugoslavia	Euro_2
Relationship	Husband	Spouse
Relationship	Wife	Spouse
Marital Status	Married-AF-spouse	Married
Marital Status	Married-civ-spouse	Married
Marital Status	Married-spouse-absent	Not-Married
Marital Status	Separated	Not-Married
Marital Status	Divorced	Not-Married
Education	10th	Dropout
Education	11th	Dropout
Education	12th	Dropout
Education	1st-4th	Dropout
Education	5th-6th	Dropout
Education	7th-8th	Dropout
Education	9th	Dropout
Education	Assoc-acdm	Associates
Education	Assoc-voc	Associates
Education	Bachelors	Bachelors
Education	Doctorate	Doctorate
Education	HS-Grad	HS-Graduate
Education	Masters	Masters
Education	Preschool	Dropout
Education	Prof-school	Prof-School
Education	Some-college	HS-Graduate