Universidad Nacional Autónoma de México

Facultad de Estudios Superiores
Acatlán

Diplomado de Ciencia de Datos
Módulo II

proyecto

Análisis de ofertas de empleo para Data Scientists

Módulo II
Profesora : Malerva Reséndiz Carla Paola
Alumno : Hernández González Ricardo Paramont
Fecha : domingo 25 de abril de 2021
Ubicación del texto (Latex) : https ://www.overleaf.com/read/pqzjqbvtpsxk

Índice

1. Presentación

1.1 Objetivo

En el presente proyecto, se presenta el procesamiento y modelado de una tabla de datos correspondiente a varias ofertas de empleo del área de ciencia de datos en los Estados Unidos, obtenidas por webscrapping del portal Glasddor. El objetivo del proyecto es generar las transformaciones de los datos y el modelo de entrenamiento supervisado adecuados para poder prodecir un rango de estimación de salario con una gran certeza.

1.2 Diccionario de datos

Variable	**Tipo**	**Descripción**
Job Title	discreto	Nombre de la oferta de trabajo.
Salary Estimate	continuo	Rango del salario estimado por el portal Glassdoor, en miles de dólares
Job Description	texto	Descripción de distintos rubros de la empresa, puesto y solicitud.
Rating	discreta	Calificación por parte de usuarios de la empresa en cuestion. Valor entre 1.0 y 5.0
Company Name	Texto	Nombre de la compañía que ofrece la oferta de trabajo.
Location	discreto	Ciudad donde se ubica el trabajo ofertado, junto con el estado o país respectivo.
Headquarters	discreto	Ciudad y estado/país donde se ubica la sede de la empresa.
Size	discreto	Divide a las empresas según su número de empleados en rangos específicos.
Founded	discreto	Año de fundación de la empresa.
Type of Owners	discreto	Tipo de compañía según razón social o identidad.
Industry	discreto	Tipo de compañía según idustria en la que trabaja.
Sector	discreto	Sector en la que la compañía ofrece sus bienes o servicios.
Revenue	discreto	Utilidades de la compañía en dolares al año.
Competitors	discreto	Principal compañía competidora de la compañía que ofrece la oferta de trabajo.
Easy Apply	discreta	Etiqueta que indica si la aplicación por el puesto es sencilla.

2. Calidad de datos

2.1 Etiquetado de variables

Se revisó el tipo de dato, valores únicos, y primeros cinco registros de cada variable para su clasificación dentro de los criterios de :

- Continua
- Categórica
- Texo
- Fecha

Los hayazgos llevaron al etiquetado de las variables de la siguiente forma :

Continuas

A pesar de que la variable objetivo es continua, se decidió mejor dividirla en categorias más delante.

Categóricas

- c_salary_estimate
- v_job$_t$*itlev_salary_estimate*
- v_rating
- v_location
- v_headquarters
- v_size
- v_founded
- v_type_of_ownership
- v_industry
- v_sector
- v_revenue
- v_competitors
- v_easy_apply
- v_salary_estimate_source'

Texo

- t_company_name
- t_job_description

Fecha

No se encontró ninguna variable de fecha.

Se observa que la gram mayoría de las variables en este problema son categóricas.

2.2 Duplicados

No se encontro ningún registro duplicado a lo largo de la tabla de datos.

2.3 Completitud

Se encontraron varias columnas con valores faltantes :

	columna	total	completitud
0	v_easy_apply	3745	4.195
1	v_competitors	2760	29.394
2	v_revenue	1392	64.390
3	v_founded	977	75.006
4	v_industry	546	86.032
5	v_sector	546	86.032
6	v_rating	409	89.537
7	v_headquarters	240	93.860
8	v_size	229	94.142
9	v_type_of_ownership	229	94.142
10	v_job_title	0	100.000
11	v_salary_estimate	0	100.000
12	t_job_description	0	100.000
13	t_company_name	0	100.000
14	v_location	0	100.000

La variable de v_easy_apply fue eliminada por contar con tantos valores faltantes, los cuales son demasiados para intentar imputarlos.
Los valores faltantes del resto de las variables, incluyendo las que contienen menos de 80% de completitud, son imputados en la sección de Valores Ausentes, tomando supuertos pertinentes.

2.4 Limpieza de texto

Se comprende como limpieza de texto a la remoción de caracteres especiales (.,-_% etc.), incluyendo caracteres acentuados y la leta 'ñ', además de la transformación de todas las mayúsculas a minúsculas. Debido a la naturaleza de algunas, se conservaron algunos caracteres especiales, como en el caso de localidad para indicar la separación ciudad, estado.
Las siguientes variables sufrieron limpieza de texto :

- v_job_title
- v_location
- v_headquarters
- v_industry
- v_sector
- v_type_of_ownership
- v_size
- t_company_name
- t_job_description

2.5 Categorización de la variable objetivo

Originalmente, el problema se había pensado para un problema de regresión donde se estimaba un valor puntual del salario de la oferta de empleo. Sin embargo, la calidad y cantidad de información complican bastante el poder obtener un modelo con buenas métricas para este objetivo. Por esto mismo, se decidió replantear el objetivo del proyecto como uno de clasificación, que estimaría el salario del puesto dentro de rangos, o categorías.

- Primero, para definir las categorías de la clase, se normalizará la columna de salario tomando el promedio entre su estimación mínima y máxima.
- Después se decide dividir el salario en tres categorias :
  1.- menor de 85K anuales - categoría 0
  2.- mayor a 85k y menor a125k anuales - categoria 1
  3.- 125k o más anuales - categoria 2

2.6 Consistencia

- Se asegura que el valor del mínimo salario estimado es mayor al salario mínimo estadounidense. Para lo mismo se considera un salario mínimo de $7.5/h, una jornada laboral de medio tiempo de 4h/día y el calendario laboral estadounidense de 2019 que fue de 261 días. No se encuentra ninguna inconsistencia.
- Se asegura que en ninguno de los registros, el salario mínimo tenga un valor mayor al salario máximo. No se encuentra ninguna inconsistencia.
- Se asegura que ninguno de los registros de la calificación de la empresa (v_rating) esté por debajo de 1 y por arriba de 5. No se encontró ninguna inconsistencia.

2.7 Normalización

Normalización de v_job_title
Una exploración simple de la variable nos permite identificar al menos 1971 valores únicos que sin embargo, comparten grandes similitudes. Por ejemplo, existen muchos puestos que son en esencia de científico de datos pero con algún nombramiento adicional como trainee, junior, senior, manager, entre otros. Por lo que se procede a agrupar grandes categorías de empleos :

- data scientist
- data engineer
- data analyst
- machine learning professional
- business intelligence analyst
- analyst of other nature
- data architect

Adicionalmente, se encuentran varios títulos que tienen una sola ocurrencia y cuyas especificaciones parecen muy especiales. A dichos empleos con una sola ocurrencia que no fueron agrupados en las categorías anteriores, se les incluyó en la categoría de highly specific. A las categorías restantes, se les agrupó en la categoría de others, la cual posee el menor número de ocurrencias.

Normalización de v_location
La variable consta de la ciudad y estado o país donde se localiza la oferta de trabajo, por lo que se extraen de ellas las variables de :

- v_city : con 191 de categorías.
- v_state : con 10 categorías.
- v_big_city : variable dummy que indica si la oferta de trabajo se ubica en una de las 20 ciudades más pobladas de Estados Unidos.

*La variable v_location es eliminada por lo tanto.

Normalización de v_headquarters
La variable consta de la ciudad y estado o país donde se localiza la sede de la empresa, por lo que se extraen de ellas las variables de :

- v_headquarters_city : consta de 522 categorías
- v_headquarters_state : 50 categorías.

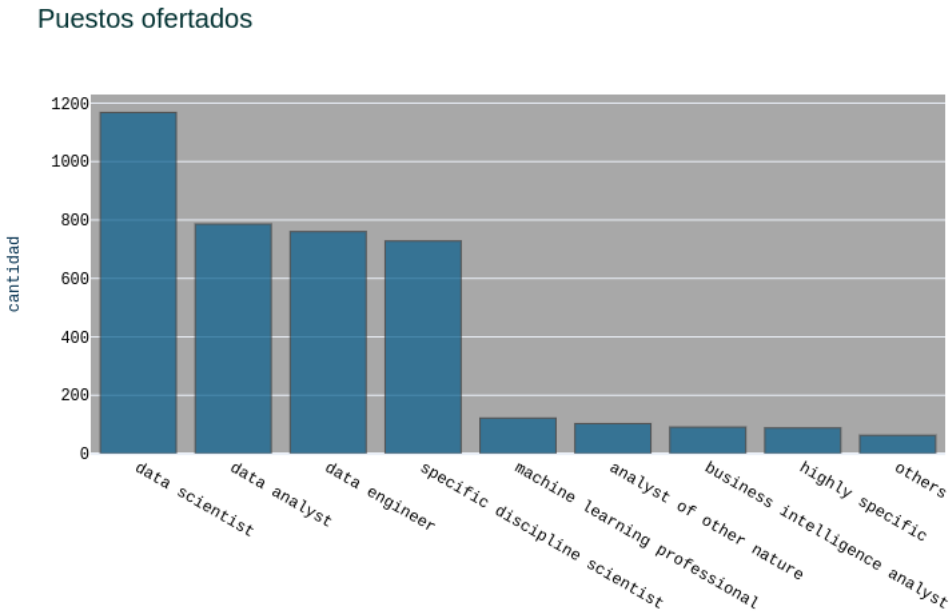
En la variable v_headquartes_state se identifica a aquellos registros de paises fuera de los Estados Unidos y se les agrupa en la categoría foreign countries.

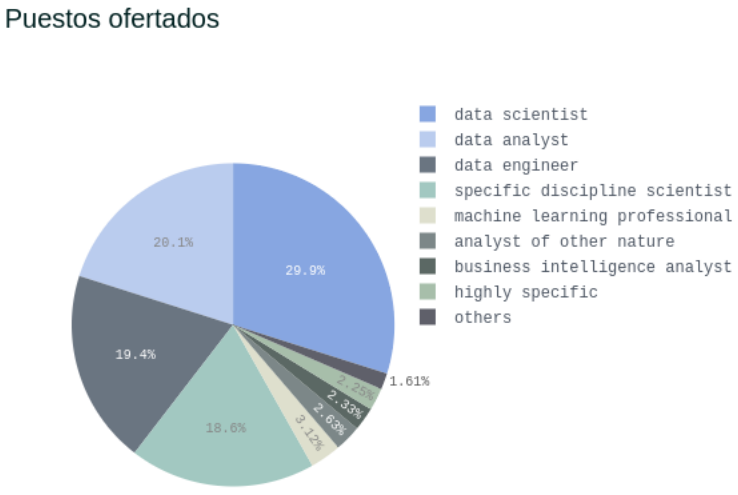
Normalización de v_industry y v_type_of_ownership

- Se agrupan todas las categorías dentro de industry con 9 o menos ocurrencias en la categoría others.
- Se agrupan todas las categorías dentro de type of ownership con 5 o menos ocurrencias en la categoría others.

3. Análisis Exploratorio de Datos

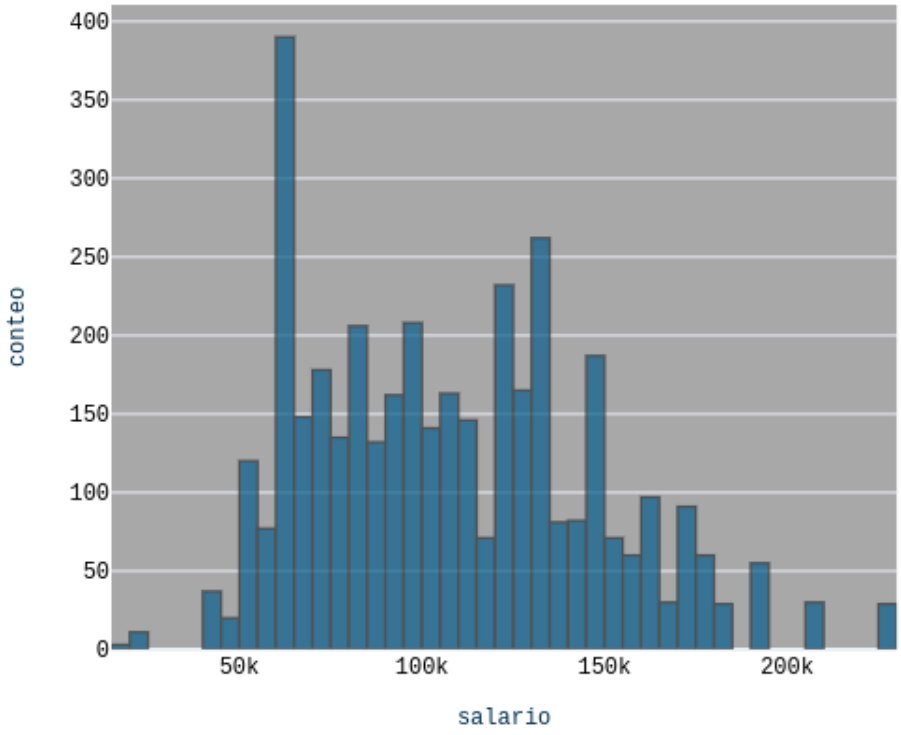
Distribución de puestos ofertados





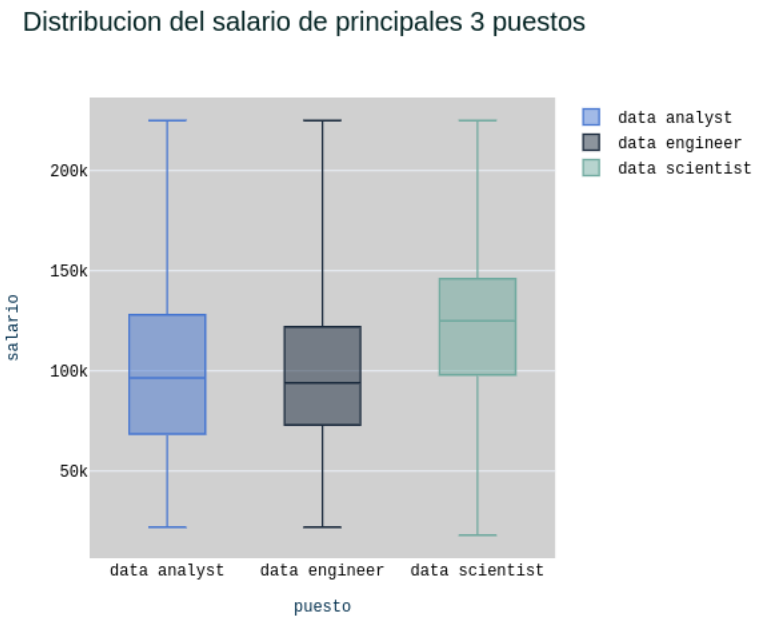
Casi uno de cada tres puestos ofertados tiene el título de Data Scientist. Los tres títulos de empleo más comunes son Data Scientist, Data Analyst y Data Engineer, entre los tres componen 70% de las ofertas.

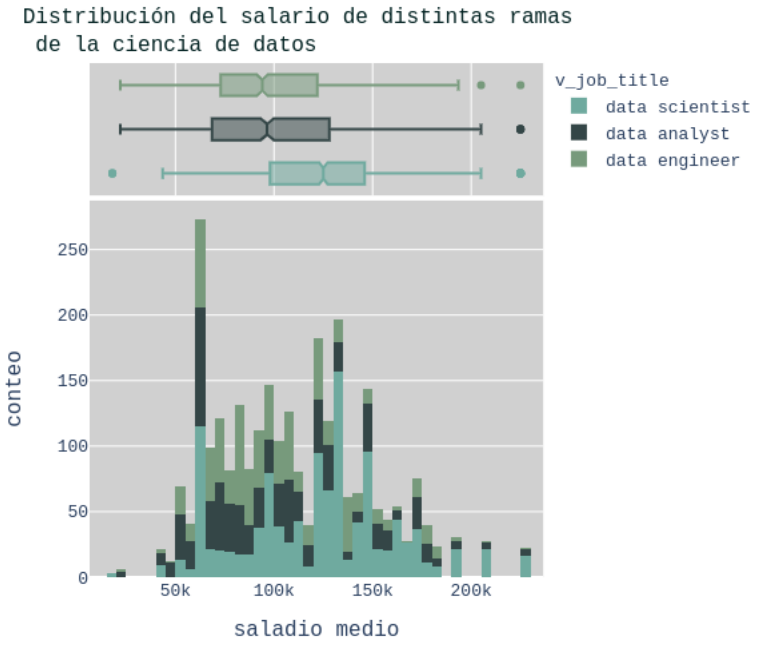
Distribución de la media de la estimación de salarios



Prácticamente todos los salarios sobrepasan los $50k anuales, la mayoría se encuentran alrededor de los $100k anuales, sin embargo, hay casos que llegan hasta los $225k anueales.

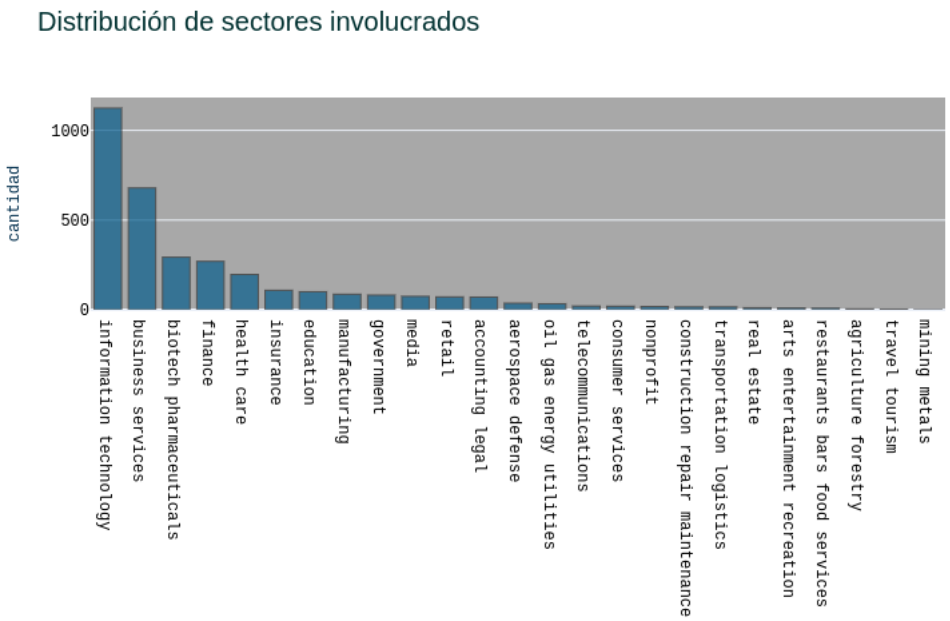
Distribución del salario de los primeros tres puestos ofertados

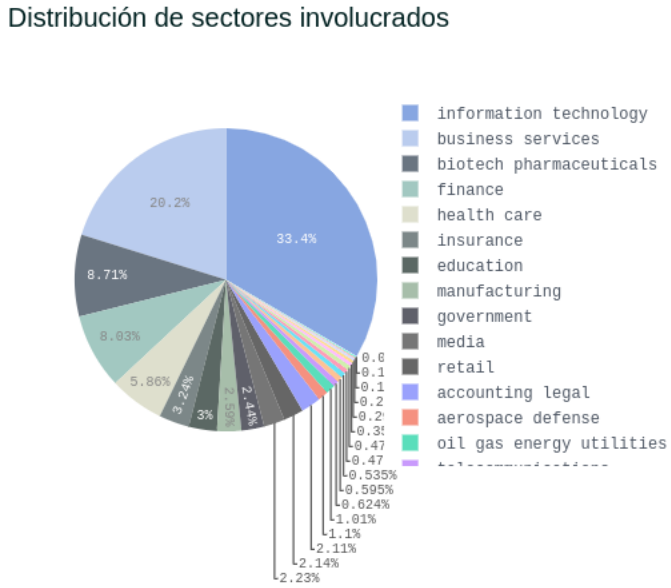




Se observa que los puestos de Data Scientist además de ser más comunes, son también mejor pagados, seguidos por los puestos de Data Analyst y Data Engineer.

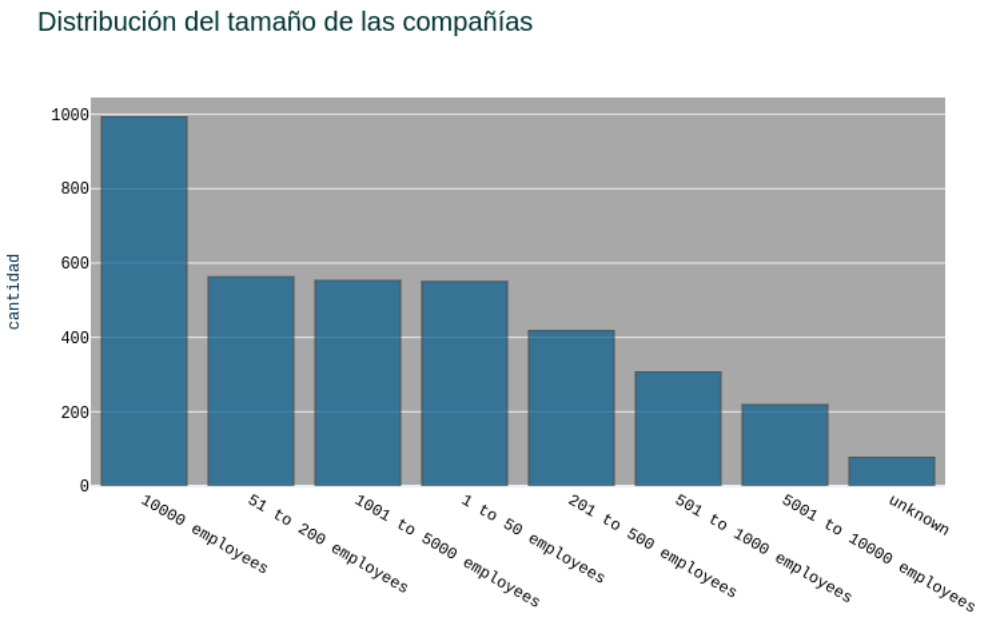
Principales sectores involucrados en las ofertas de trabajo

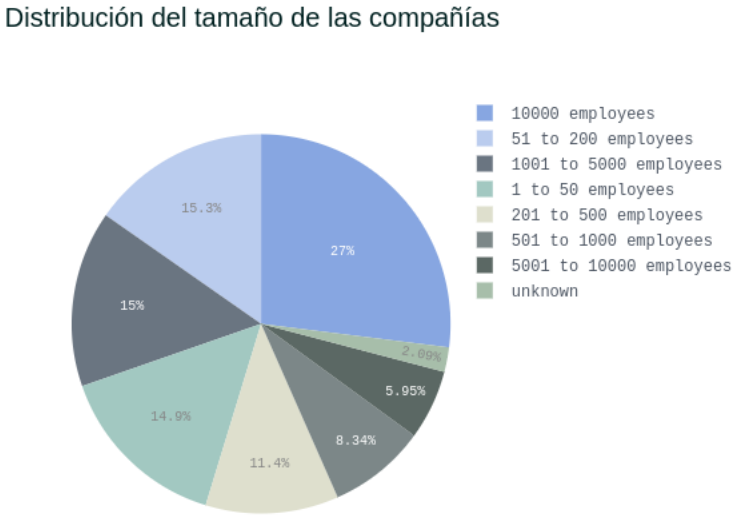




Las tecnologías de la información constituyen el principal sector involucrado en las ofertas de empleo. Uno de cada tres empleos es ofertado dentro del mismo. El sector de business services es el siguiente más común, la mitad de las ofertas de empleo son acaparadas entre estos dos.

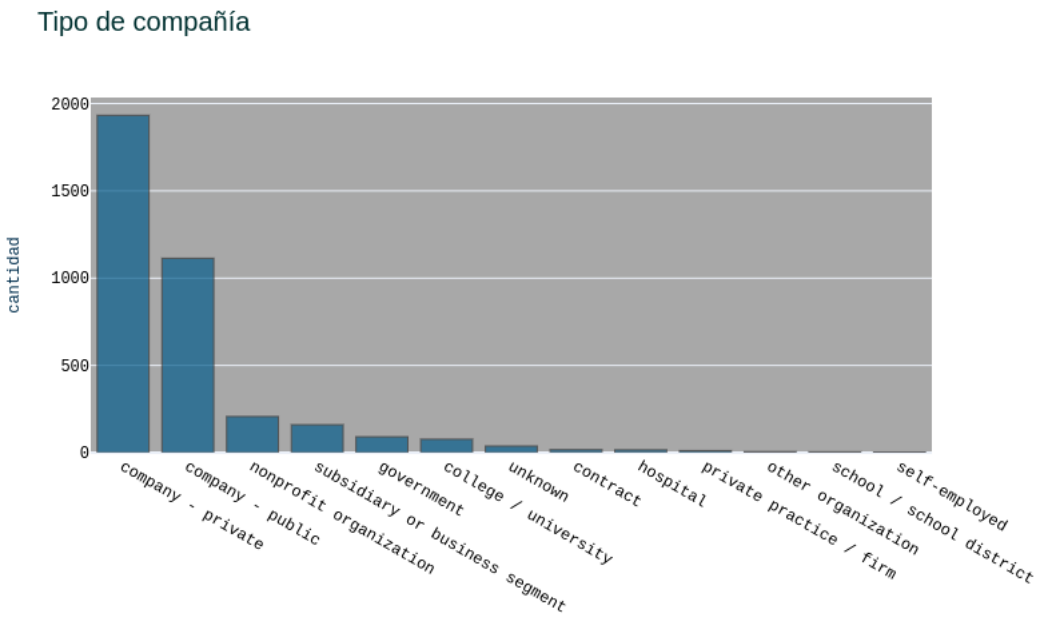
Distribución del tamaño de las compañías ofertando empleo

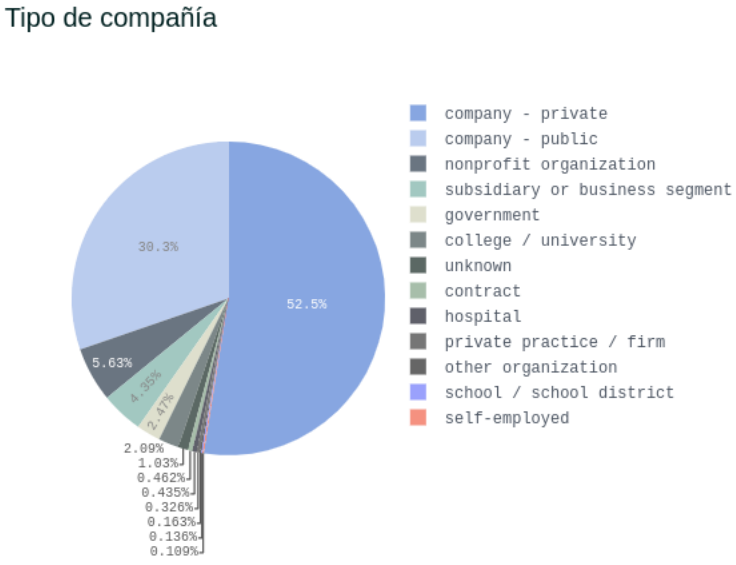




Las mayoría de los empleos son ofertados por las compañías más grandes dentro de la clasificación de la tabla (1000 o más emleados). Sin embargo, son empresas pequeñas (entre 51 y 200 empleados) las segundas en ofertar más empleos.

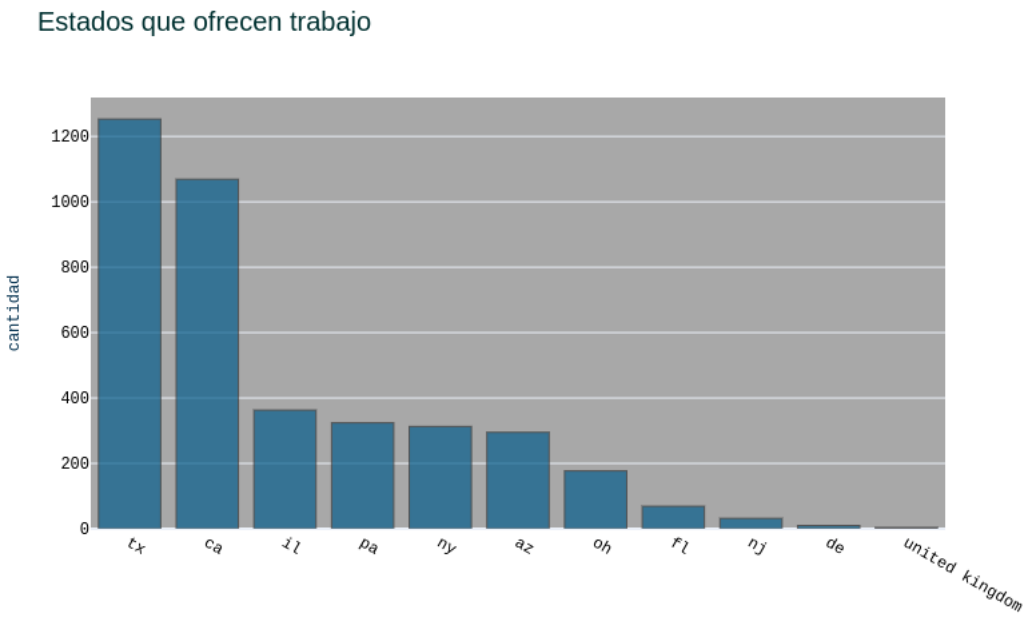
Distribución del tipo de propiedad de las compañías

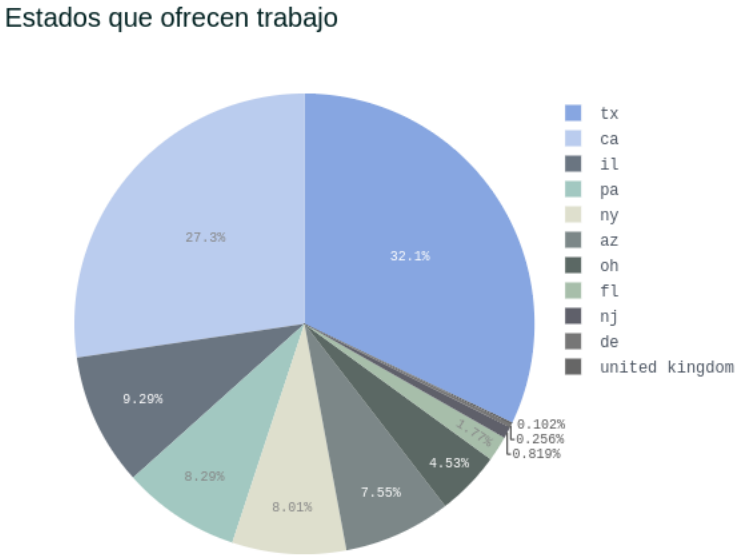




Más de la mitad de las compañías son privadas y aproximadamente un tercio son públicas. Apenas 0.1% de los trabajos ofertados se identifican dentro del autoempleo.

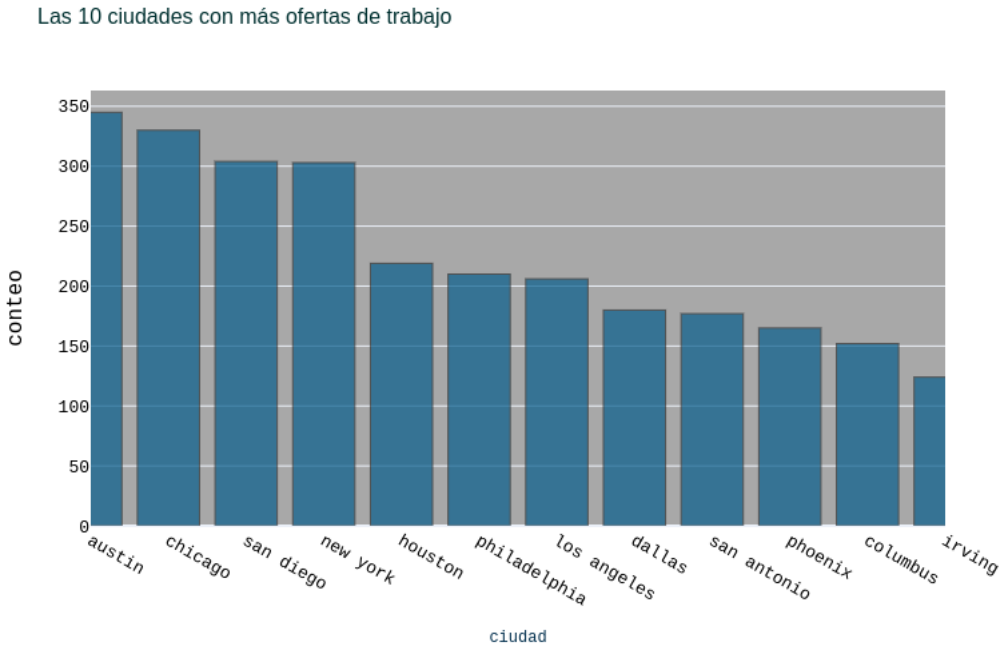
Estados donde se ubican los empleos ofertados





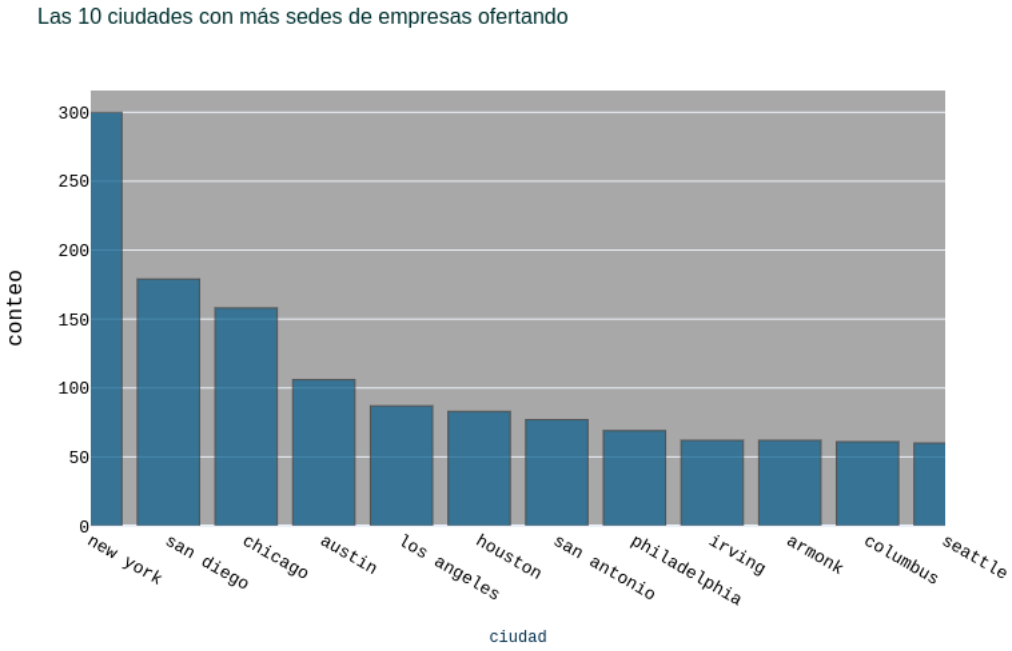
La mayoría de los empleos ofrecidos se ubican en Texas, fruto de la transición de la economía del estado del sector energético al sector de las tecnologías de la información. No muy lejos, sigue California, que representa el estado lider en inovaciones de tecnologías digitales y el que más aporta al PIB nacional. Curiosamente, se observa que algunos de los empleos ofertados son en el Reino Unido.

Las 10 ciudades con más ofertas de empleo



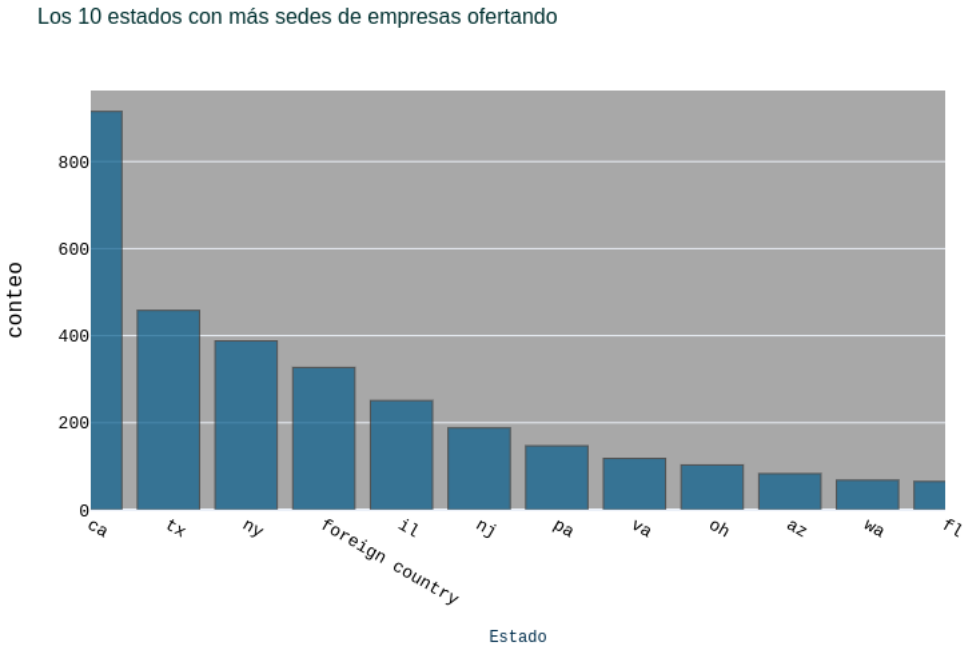
La ciudad que más empleos oferta es Austin, Tx, la cual fue calificada como la segunda con mayor crecimiento económico en Estados Unidos en 2020, superada sólo por Denver, Tx. La seguda ciudad que más empleos oferta es Chicago, la cual representa la economía más grande fuera de las costas estadounidenses.

Las 10 ciudades con más sedes (headquarters) de compañías que ofertan empleos



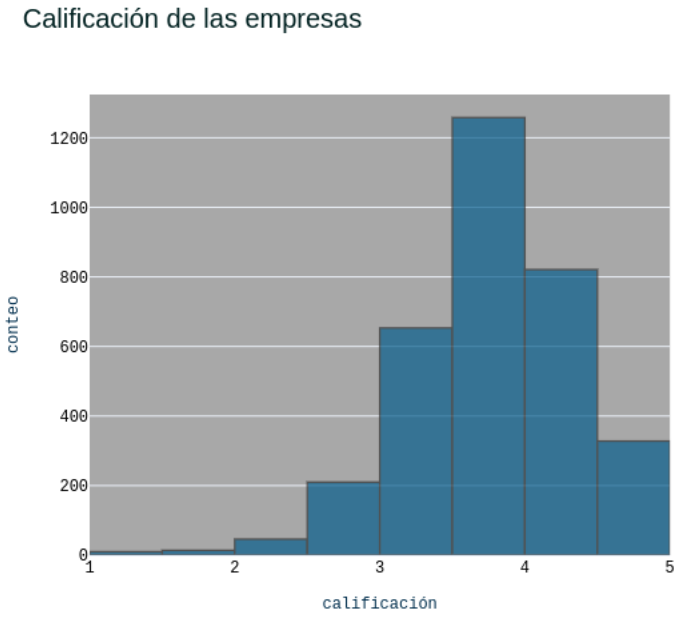
La ciudad que aloja más sedes de las compañías involucradas es Nueva York, superando a San Diego, el segundo lugar, por casi el duble. Curiosamente, la única ciudad de california en el top 10 es Los Angeles, apenas en el puesto 5, a pesar de que California es el estado lider en tecnologías digitales.

Los 10 estados con más sedes de empresas ofertando



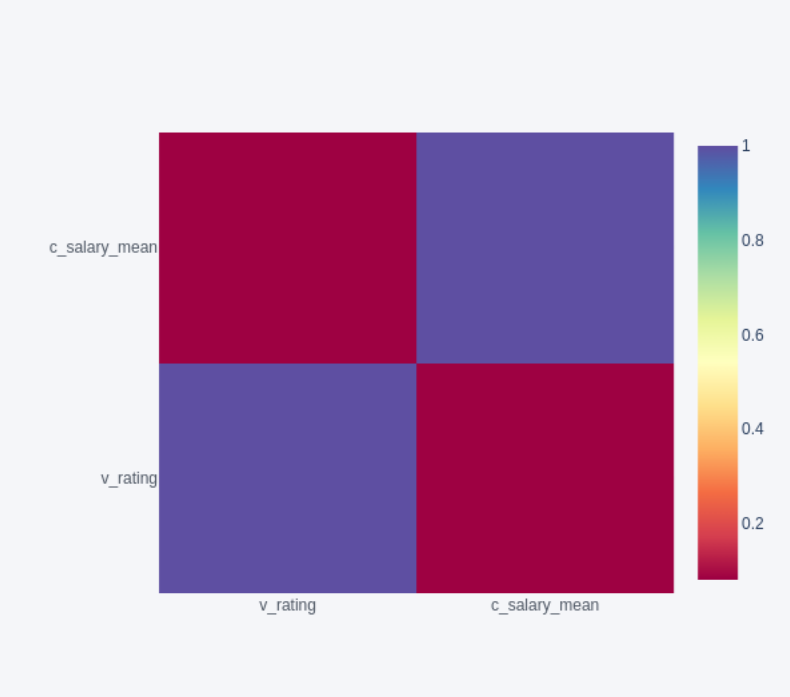
A pesar de que sólo una ciudad de California se encuentra dentro del top 10 de ciudades con más sedes, California concentra la mayoría de sedes a lo largo de sus ciudades.

Distribución de la calificación de las empresas



Se nota que la tendencia es que las empresas sean calificadas con una calificación de 3.5 estrellas en una escala de 1 estrella a 5 estrellas.

Correlación entre calificación de las empresas y salarios que ofrecen



Se aprecia que no existe una correlación significativa entre la calificación de la empresa y los salarios que ofertan. Lo que indica que los empleados no toman en cuenta el salario entre los criterior principales para calificar su entorno de trabajo.

Palabras más mencionadas en la descripción de las ofertas de empleo

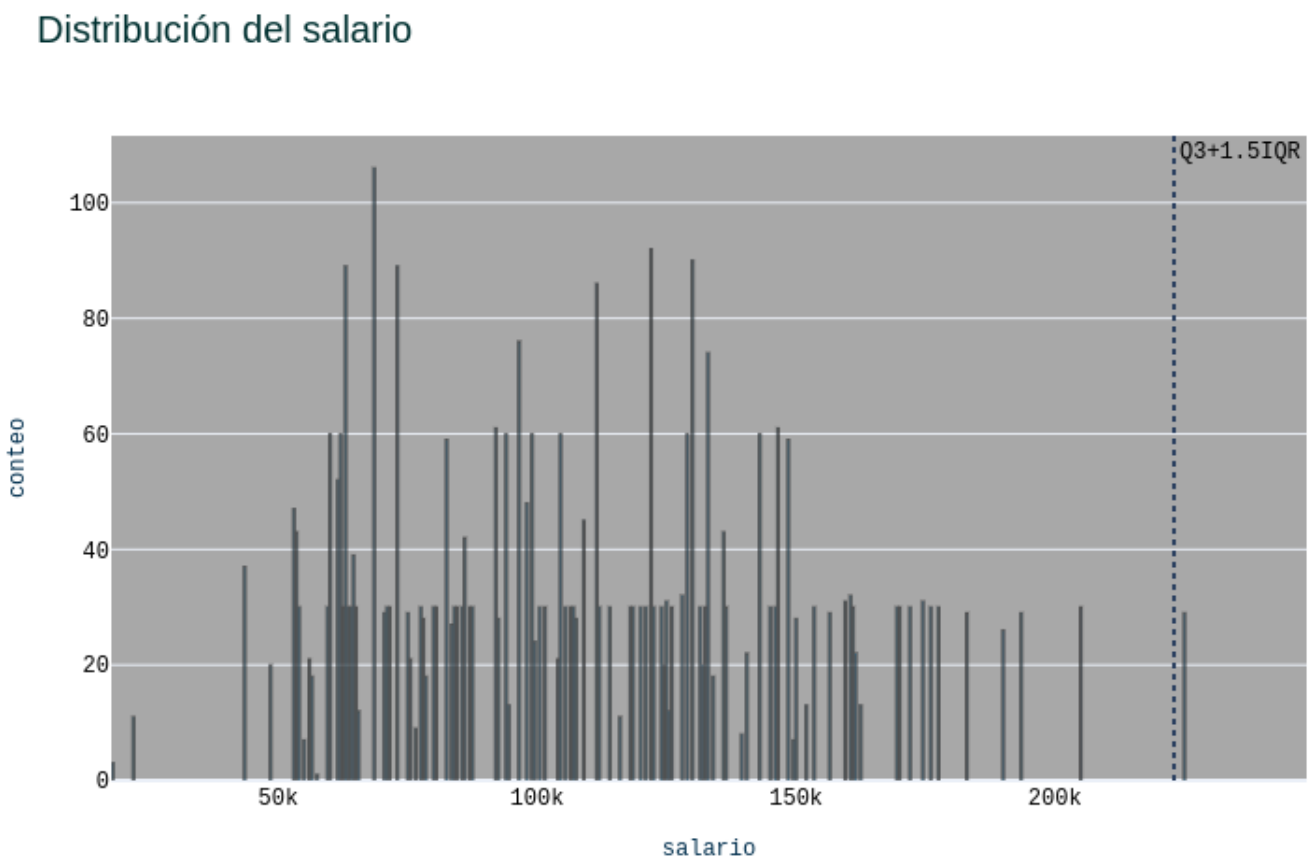
La palabra más común es 'Data', lo que resulta obvio. Las siguientes más mencionadas según su orden son experiencia, negocios y habilidades. Si nos enfocamos en lenguajes de programación, los más mencionados son python, seguido por R y después java.

Palabras más comunes en el nombre de las empresas

Se alcanza a apreciar que los empleos ofertados por Apple, Amazon, IBM, Facebook e incluso la Universidad de Texas son suficientes para resaltar en el gráfico.

4. Valores anómalos

Se decide hacer un analisis de valores atípicos para la variable objetivo de salario. Se encuentra que existen 29 valores atípicos demasiado altos (0.74% de los registros totales), que sobrepasan el límite heurístico de distribución de Q3 + 1.5IQR. Sin emabrgo, observando dichos valores, se descubre que estos sobrepasan el límite heurístico de distribución por apenas 1.85% el valor de la media del salario.



Este hecho, junto con que dichos salarios tan altos son de gran interés para su inclusión en el modelo predictivo hacen tomar la decisión de conservarlos.

5. Valores ausentes

Se cuenta con 10 variables que contienen valores ausentes a imputar :

	columna	total	completitud
0	v_competitors	2213	29.229
1	v_revenue	1090	65.142
2	v_founded	777	75.152
3	v_industry	427	86.345
4	v_sector	427	86.345
5	v_rating	320	89.767
6	v_headquarters_state	197	93.700
7	v_headquarters_city	195	93.764
8	v_size	184	94.116
9	v_type_of_ownership	184	94.116

- Se imputan con la media las variable de v_raiting.
- Se imputan con la moda las variables de :
  - v_type_of_ownership
  - v_size
  - v_headquarters_city
  - v_headquarters_state
  - v_sector
  - v_industry
  - v_founded

- La variable v_competitors se transforma en una dummy, donde el 1 significa que la empresa tiene competidores. Adicionalmente se agrega la variable v_competitors_count que tiene el número de competidores de la variable original.
- Se codifica la variable v_revenue como una categórica ordinal, tomando los valores faltantes como que la empresa gana demasiado poco, por lo tanto los valores faltantes equivalen a la categoría de menor orden. La categorización se hace en la sección de ingeniería de variables.

6. Ingeniería de variables

6.1 Nuevas variables

- Se incluye una columna que revisará si la compañia (t_company_name) está dentro de las empresas FAANG o Big Tech. Se consideran las empresas :
  - facebook
  - apple
  - netflix
  - google
  - microsoft
  - ibm
  - spotify
  - tesla
  - uber
  - twitter
  - alphabet
  - visa
  - envidia
  - intel
  - adobe
  - cisco
  - att
  - oracle
  - airbnb
  - samsung
  - foxconn
  - huawei
  - dell
  - sony
  - hp
  - lg
  - lenovo
- Se incluye variable dummy v_is_bigcity que indica si el registro se encuentra entre las 20 ciudades más pobladas de Estados Unidos : • 'New York'
  - 'Los Angeles'
  - 'Chicago'
  - 'Houston'
  - 'Phoenix'
  - 'Philadelphia'
  - 'San Antonio'
  - 'San Diego'
  - 'Dallas'
  - 'San Jose'
  - 'Austin'
  - 'Jacksonville'
  - 'Fort Worth'
  - 'Columbus'
  - 'Charlotte'

- 'San Francisco'
- 'Indianapolis'
- 'Seattle'
- 'Denver'
- 'Washington'

6.2 One-Hot encoding (Dummies)

Se decide codificar a las variables categóricas sin orden específico con el uso de dumificación. Dichas columnas son :

- v_job_title
- v_type_of_ownership
- v_industry
- v_sector
- v_city
- v_state
- v_headquarters_city
- v_headquarters_state

6.3 Ordinal-Encoding

Para las variables categoricas ordinales v_size y v_revenue, se codifican sus valores especificando el orden. A la categoría de menor orden se le asigna el valor 0 en adelante. El orden es :

- v_size
  'unknown'
  '1 to 50 employees'
  '51 to 200 employees'
  '201 to 500 employees'
  '501 to 1000 employees'
  '1001 to 5000 employees'
  '5001 to 10000 employees'
  '10000 employees'
- v_revenue
  'unknown',
  'Less than $1 million (USD)'
  '$1 to $5 million (USD)'
  '$5 to $10 million (USD)'
  '$10 to $25 million (USD)'
  '$25 to $50 million (USD)'
  '$50 to $100 million (USD)'
  '$100 to $500 million (USD)'
  '$500 million to $1 billion (USD)'
  '$1 to $2 billion (USD)'
  '$2 to $5 billion (USD)'
  '$5 to $10 billion (USD)'
  '$10+ billion (USD)'

6.4 Texto

- A las variables de texto (t_job_description y t_company_name) se les extrae la longitud de texto, contando número de caracteres y número de palabras (las palabras se cuentan por separación con espacios blancos). Se añaden las variables correspondientes.
- Se le aplica vectorización a la columna t_job_description con la tecnica de coun vectorizing, con la condición de que la palabra necesitaba una ocurrencia de al menos 5

7. Reducción de dimensiones

7.1 Filtro de baja correlación con el objetivo

Con el objetivo de eliminar variables de entrada, o independientes, que pueden ser muy poco relevantes, se evalúa la correlación de spearman entre las columnas independientes y el objetivo. Aquellas con un valor menor a 0.001 son eliminadas.

7.2 Multicolinealidad

Se estudia la multicolinealidad entre las variables generadas por la dumificación y la vectorización del texto y se decide eliminar a las altamente relacionadas con el criterio de que su VIF (variable inflation factor) es mayor a 10.

8. Modelado

Se entrenan distintos modelos de un método de predicción específico, con el uso de grid o random search y cross validation. Se elige el mejor modelo del método en cuestión según su valor de exactitud obtenido por cross validation. Al final, se eligen los mejores métodos comparando sus méticas generalizadas, con especial enfásis en el roc-auc de cada una de las tres clases para los datos de prueba, obtenido con el método One vs Rest. A continuación, se muestran las méticas de los tres mejores modelos encontrados.

Proporción entre conjunto de entrenamiento y prueba

	Resgitros	**Porcentaje**
Train	3127	80%
Test	782	80%

Tasa de evento entre en el conjunto de entrenamineto

Clase	**Registros**	**Porcentaje**
0	1054	34%
1	1010	34%
2	1063	32%

Tasa de evento entre en el conjunto de prueba

Clase	**Registros**	**Porcentaje**
0	264	34%
1	252	34%
2	266	32%

Codificación de la variable objetivo : Estimación del Salario

Clase	Codificación
0	Menor a $85k anuales
1	Entre $85k y $125k
2	Mayor a $125k

Modelos probados

Modelo	exactitud train	exactitud test
KNN	0.538	0.510
Reg Logistica	0.904	0.503
SVC	0.670	0.568
Arboles Dec	0.584	0.594
Ran Forest	0.637	0.564
SGD	0.618	0.613
GB	0.767	0.597
Ada Boost	0.577	0.581

8.2 Gradiente Estocástico Descendente

Gradiente estocástico descendente compone el mejor modelo encontrado para la multilasificación. Posee las mejores métricas de exactitud (accuracy) y de roc-auc para los datos de test.

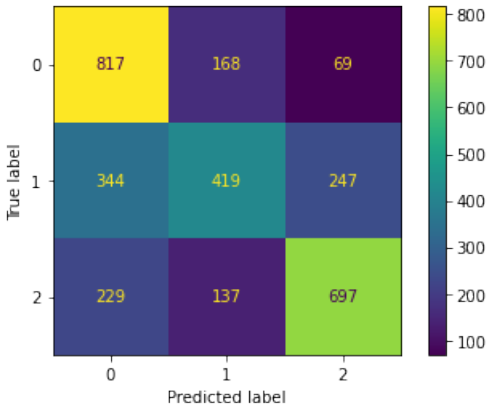
Metricas

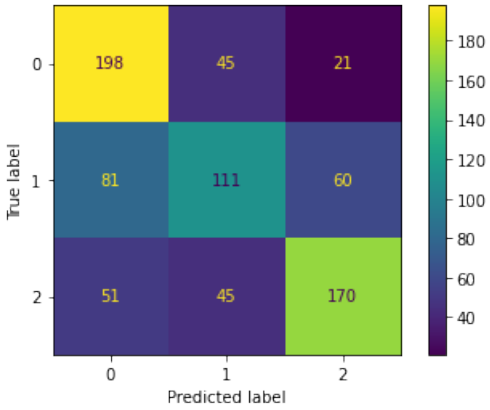
Métricas Train

	precision	recall	f1-score	support
0	0.588	0.775	0.669	1054
1	0.579	0.415	0.483	1010
2	0.688	0.656	0.671	1063
accuracy			0.618	3127
macro avg	0.618	0.615	0.608	3127
weighted avg	0.619	0.618	0.610	3127

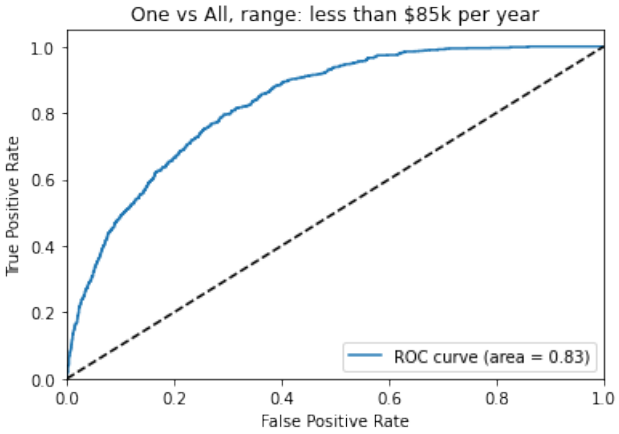
Métricas Test

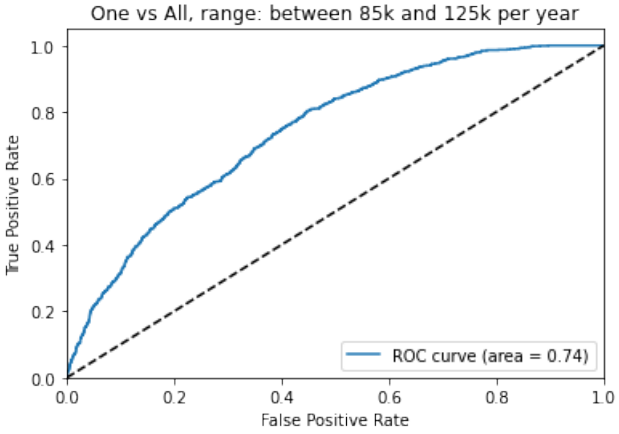
	precision	recall	f1-score	support
0	0.600	0.750	0.667	264
1	0.552	0.440	0.490	252
2	0.677	0.639	0.658	266
accuracy			0.613	782
macro avg	0.610	0.610	0.605	782
weighted avg	0.611	0.613	0.607	782

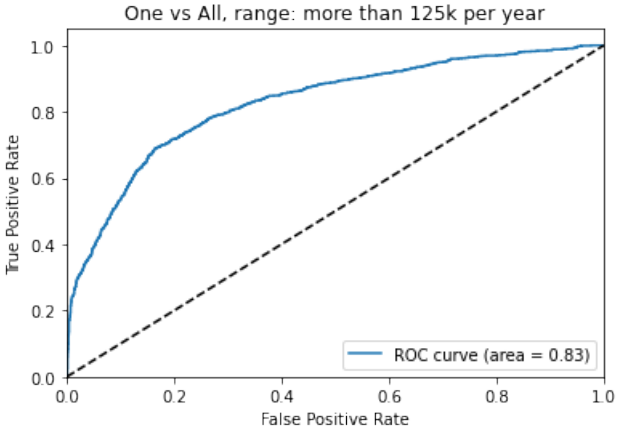

817
168
69
344
419
247
229
137
697
True label
Predicted label
0
1
2
800
700
600
500
400
300
200
100


198
45
21
81
111
60
51
45
170
True label
Predicted label
0
1
2
180
160
140
120
100
80
60
40

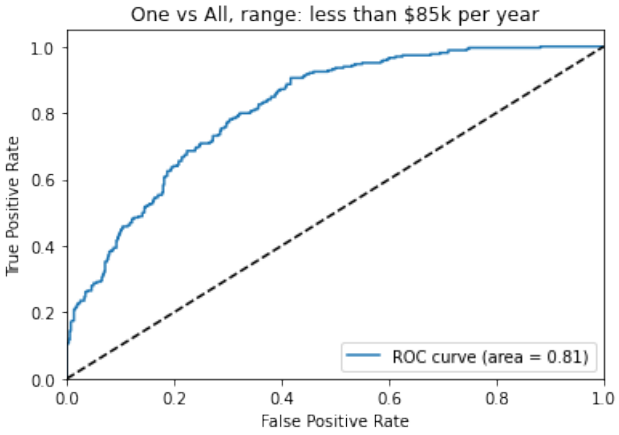
Curvas roc con los datos de entrenamiento

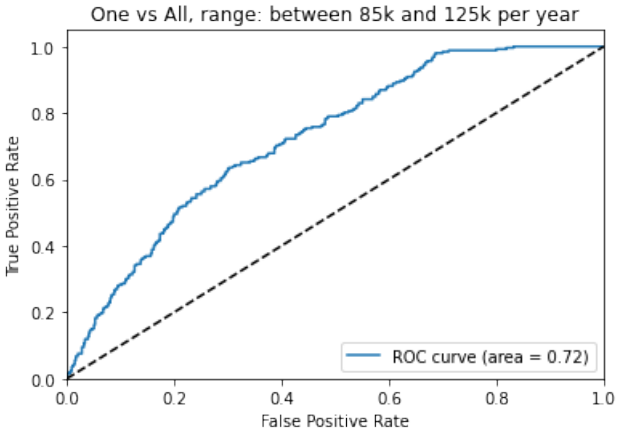


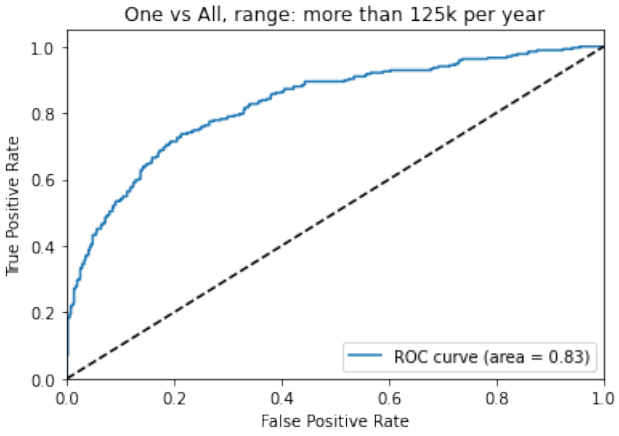




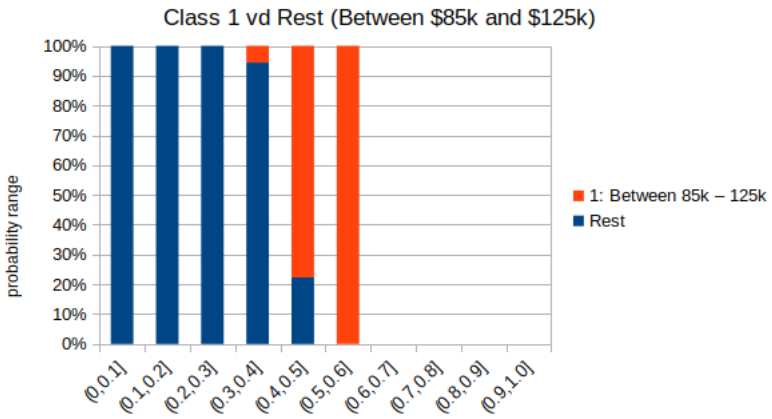
Curvas roc con los datos de prueba

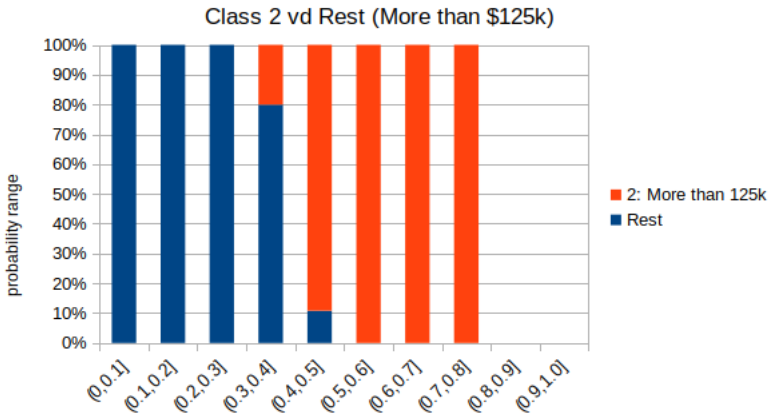






Estabilidad





Se observa que las preicciones del modelo tienen una certeza de al menos un 30% en todas la categorías de la multiclasificación. Sin embargo, el modelo no cuenta con grandes probilidades de certeza, el máximo es 80% para las clases 0 y 2, sólo 60% para la clase 1. Sin embargo, es muy buen indicio que la clase de interés de cada caso tiene la mejor distribución de probabilidad frente al resto de las clases.

8.2 Gradient Boost

Gradient Boost compone el segundo mejor modelo encontrado para la multilasificación. Se incuyen las métricas de este modelo para poder hacer una comparación simple con el mejor modelo.

Metricas

Métricas Train

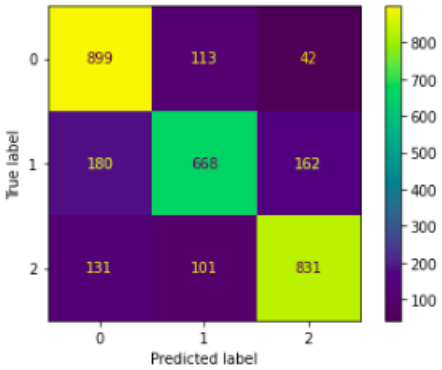
	precision	recall	f1-score	support
0	0.743	0.853	0.794	1054
1	0.757	0.661	0.706	1010
2	0.803	0.782	0.792	1063
accuracy			0.767	3127
macro avg	0.768	0.765	0.764	3127
weighted avg	0.768	0.767	0.765	3127

Métricas Test

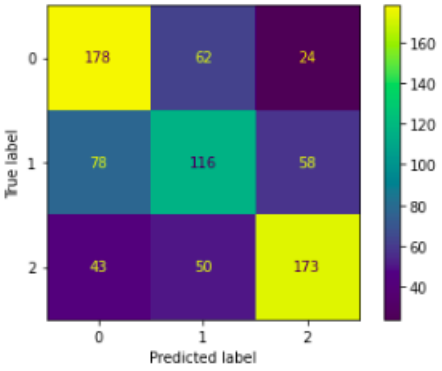
	precision	recall	f1-score	support
0	0.595	0.674	0.632	264
1	0.509	0.460	0.483	252
2	0.678	0.650	0.664	266
accuracy			0.597	782
macro avg	0.594	0.595	0.593	782
weighted avg	0.596	0.597	0.595	782

Se observa que a pesar de que las métricas para el conjunto de entrenamiento son mejores comparadas a las del modelo de gradiente estocástico descendente, las métricas de prueba no lo son. La brecha entre las métricas de entrenamiento y prueba para este modelo señalan un ligero sobrentrenamiento.

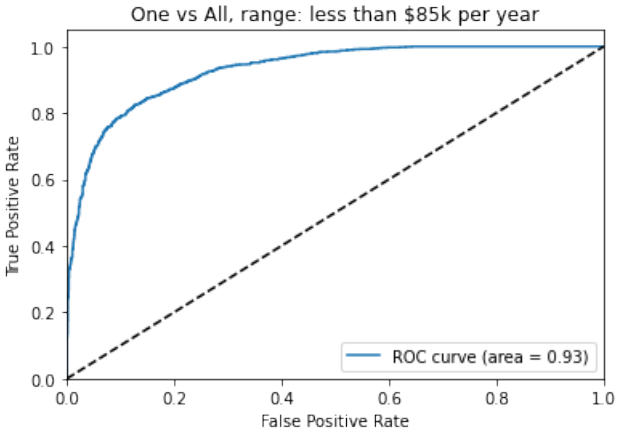
Matríz de confusión de Train

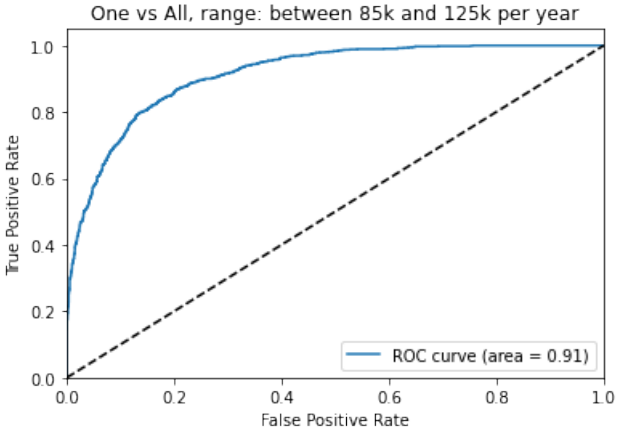


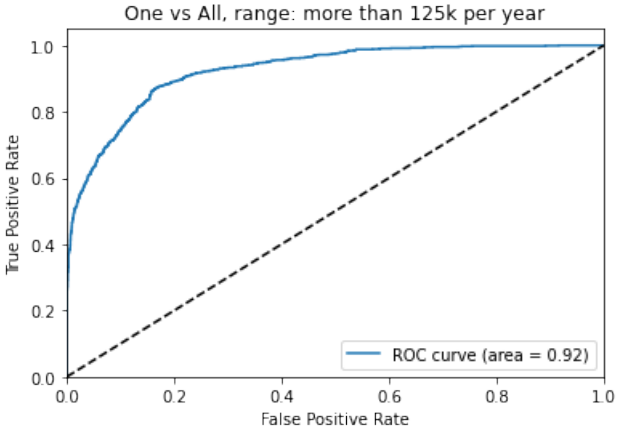
Matríz de confusión de Test



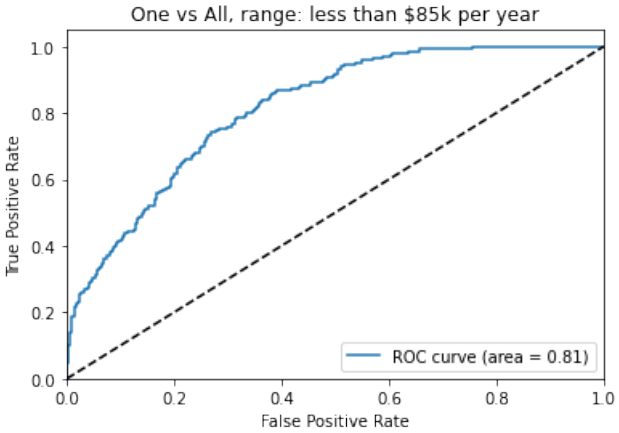
Curvas roc con los datos de entrenamiento

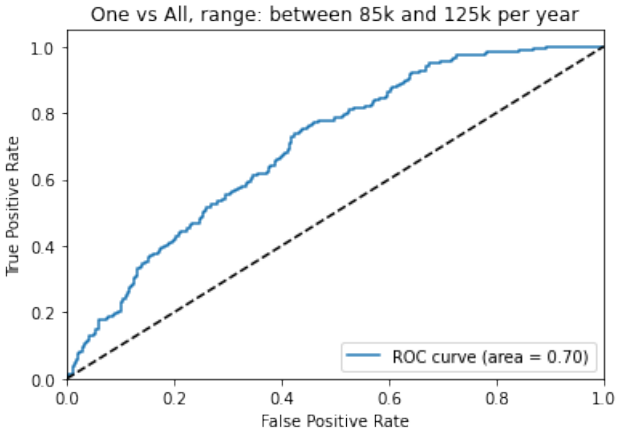


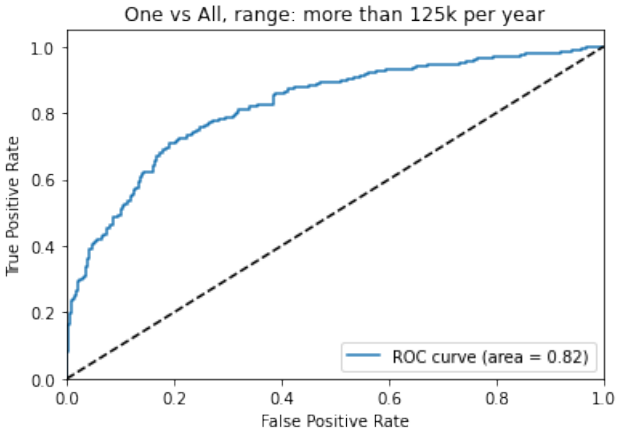




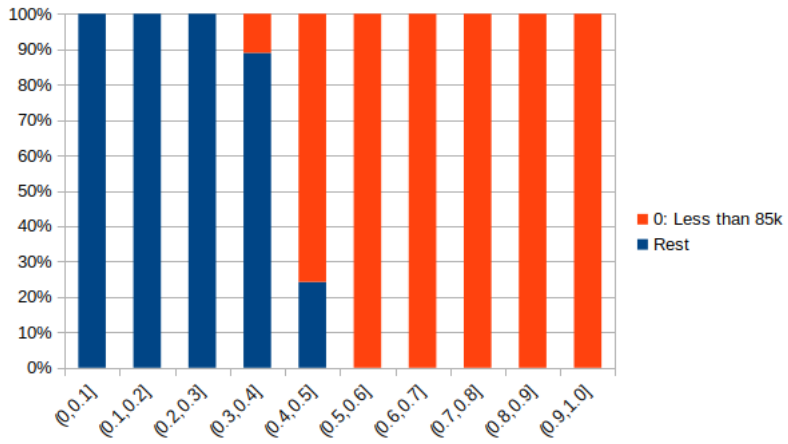
Curvas roc con los datos de prueba

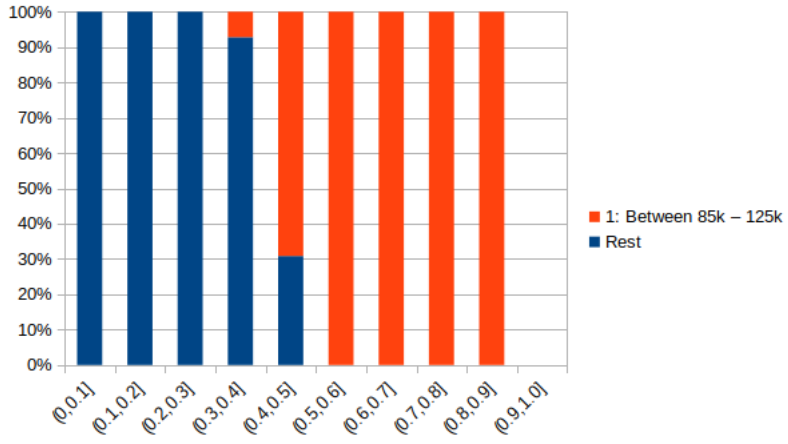


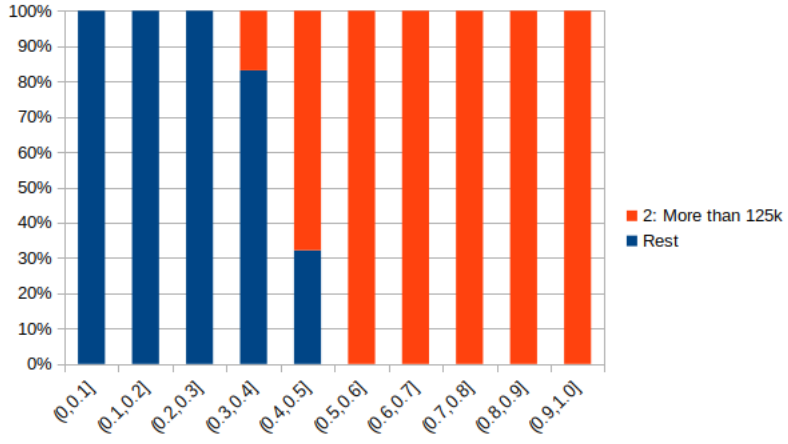




Estabilidad







A pesar de que este modelo se considera como el segundo mejor de entre los probados, sí posee una mejor distribución de probabilidad para sus predicciones.