Aplicación de Modelos de Machine Learning para predecir empleo

Luis Antonio Bermeo Fernández

Octubre 2020

1. Machine Learning

– predicción del empleo (asegurados totales IMSS) para el mes de octubre 2020 (publica en noviembre), Usar como maximo desde el periodo 2015.

1.1. Empleo

Para realizar la predicción usamos la fuente de datos abiertos del imss para obtener los datos de asegurados por mes y año desde 2015 la cuál utilizamos como variable proxy del nivel de empleo a nivel nacional. Adiconalmente de la misma fueste se obtuvieron otras variables que se piensan pueden ser explicativas del nivel de empleo como lo son el promedio de la masa salarial por sector (asi como para trabajadores eventuales y por región rural o urbana), el numero de puestos de trabajos en zonas urbanas y rurales, puestos eventuales totales y con salario asociado.
Se tomó en cuenta la variable de sector económico, así como los asegurados no trabajadores y el promedio de asegurados que eran mujeres (pues se puede pensar que si existe una discrimación laboral que les haga conseguir un trabajo con mayor dificultad que los varones entonces, para los periodos con menor porcentaje de mujeres el crecimiento del empleo podría ser mas lento).
Como variables adicionales que se consultarón en el INEGI y que pudieran ser explicaciones de los niveles de empleo se encuentran la inflación anual, el indice global de remuneraciones, variacion global del indice de remuneraciones, indice de confianza al consumidor, variacion mensual del indice de confianza del consumidor, tasa de desocupación, variación mensual de las exportaciones y variación mensual de las importaciones.
Para determinar si las observaciones del pasado influyen en las observaciones futuras se ocupo la variable de interés (asegurados) rezagada 1 y 2 periodos (meses) como variables explicativas.

Por otro lado, dado que nos interesa realizar una predicción sobre el nivel de empleo del siguiente periodo y dado que no contamos con información previsible sobre las variables dependientes ocupamos las variables explicativas rezagadas con el fin de tener una observación de variables explicativas que nos permitan realizar la predicción futura. A continuación se muestra una lista con las variables ocupadas y el número de rezagos realizados, así como una descripción.

Tabla 1: Descripción de variables

Variable	rezago	Descripción
asegurados	NA	Total de trabajadores por mes
i_a1	1	Inflación Anual
e_vm2	2	variación mensual de las exportaciones
i_vm2	2	Variación mensual de las importaciones
desocup7	7	Tasa de desocupación
igrese3	3	Índice Global de remuneraciones
ingrese_vm7	7	Variación mensual del índice global de remuneraciones
icc1	1	Índice de confianza al comsumidor
icc_vm2	2	Variación mensual del ICC
icem1	1	Índice de consumo mensual
masa_sal$_t a1$	1	Masa salarial puestos de trabajo afiliados
masa_sal_teu1	1	Masa salarial puestos de trabajo eventuales urbanos
masa_sal_tec1	1	Masa salarial puestos eventuales campo
masa_sal_tpu1	1	masa salarial permanentes urbanos
masa_sal_tpc1	1	masa salarial permanentes campo
mujer1	1	porcentaje de mujeres
no_trabajadores1	1	asegurados no trabajadores
ta1	1	puestos de trabajo afiliados
teu1	1	eventuales urbanos
tec1	1	eventuales campo
tpu1	1	permanentes urbanos
tpc1	1	permanentes campo
sector_economico_11	1	trabajadores en sector primario
sector_economico_21	1	trabajadores en sector secundario
sector_economico_41	1	trabajadores en sector terciario
asegurados_1	1	total de asegurados rezagada un mes
asegurados_2	2	total de asegurados rezagada dos meses

Para realizar la predicción se ocuparon 3 modelos de machine learning: LASSO, PCR y Random Forest los cuales se presentan a continuación en este orden.

1.1.1. LASSO

Dado que nuestra base de datos contiene una gran cantidad de predictores nos gustaría ocupar métodos un método que nos permita identificar cuales son las variables significativas para el modelo, por esta razón rechazamos el uso de Rige regression (pues en este enfoque todos los coeficientes son significativos) y en su lugar escogemos el modelo LASSO para realizar la selección de variables.
Lo primero a realizar es la calibración del modelo utilizando cross-validation, es decir la elección de λ óptimo tal que mínimize los MSE, corremos la función lasso para diferentes valores de penalización (en un rango de 0.01 a 10^{10}). La siguiente figura muestra los coeficientes estandarizados versus el logaritmo de los diferentes valores de lamda, en ella se puede ver la rapidez con la que la mayoria de los coeficientes convergen a cero, lo cual indica que muchas de nuestras variables no son explicativas del nivel de empleo.

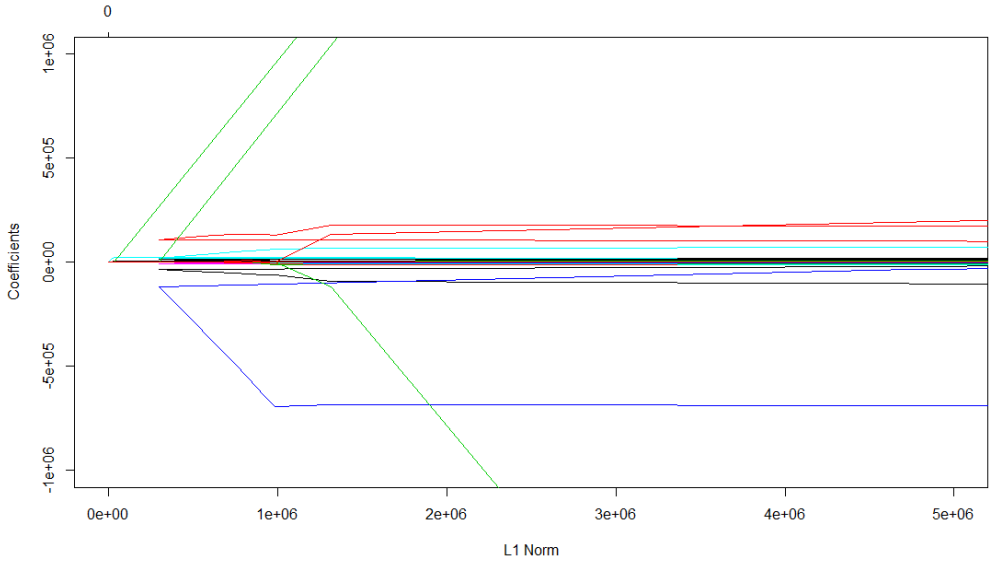

Figura 1: coeficientes vs $ln(\lambda)$

Ahora se realiza la selcción de λ^* (óptimo) como que minimize el MSE (mean square error) (esto se realiza con una sub muestra tomada aleatoriamente de la muestra original por lo que en realidad se minimiza el MSE de la submuestra). En la figura 2 se muestra el valor del MSE para diferentes valores del $log(\lambda)$.

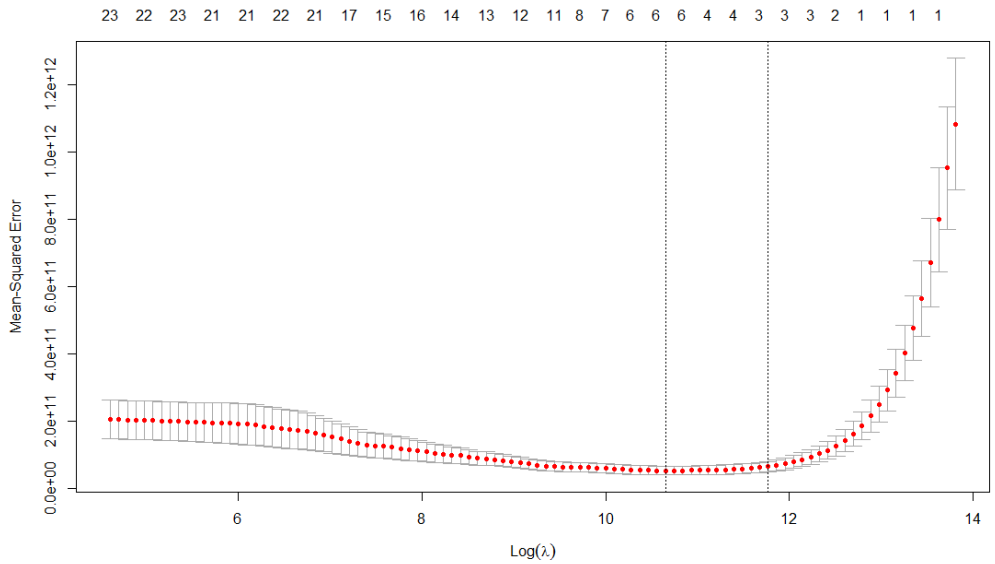

Figura 2: MSE para diferentes valores del parámetro de penalización

El nivel de $\lambda^* = 42276.92$ con este valor óptimo se volvió a correr el método LASSO y se calculo el estimado del test MSE para su posterior comparación con los otros modelos.
Para el nivel óptimo de la variable de penalización se obtuvieron las siguientes variables no zero del modelo.
Recordando la tabla descriptora la variable rezagada de la variación mensual de las remuneraciones influye negativamente

Variable	coefficient
Intercept	3203761.71
ingrese_vm7	-17.95
icc1	11596.10
tpu1	0.26
asegurados_1	0.71

al nivel de empleo esto puede deberse a que en periodos donde las remuneraciones son mas volatiles son preambulo a crisis

económicas que disminuyan los niveles de empleo, el indice de confienza al consumidor parece ser la variable mas importante (por el tamaño de su coeficiente) y esto es resultado de la estrecha relación entre consumo y producción; y producción y trabajo. La variable de trabajadores permanentes urbanos tiene un efecto positivo pequeño y puede deberse a que los cambios en la demanda laboral son mas elasticos en las zonas urbanas que en las zonas rurales (veáse que solo se tiene trabajadores permanente por lo que los trabajadores eventuales no son útiles bajo este modelo para predecir el nievl de empleo, lo cual parece bastante razonable). Por último el nivel de asegurados rezagado por un periodo fue el único significativo por lo que se puede intuir que sólo el pasado inmediato influye en el futuro.

Con este enfoque se puede realizar una predicción del nivel de empleo para el siguiente periodo, en la figura siguiente se muestran los datos del nivel de empleados (asegurados) que se tienen en la base de datos (puntos) y la predicción ajustada para cada periodo (linea continua) y veáse que esta linea tiene una observación más correspondiente a la predicción del siguiente periodo.

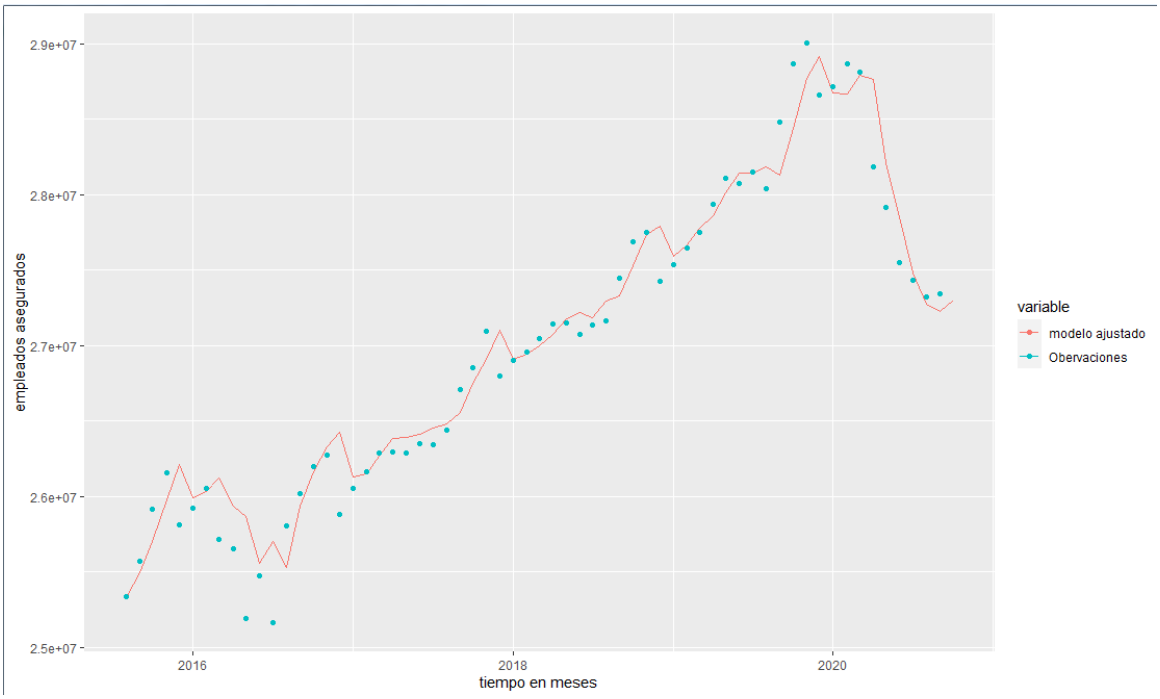

Figura 3: modelo LASSO

En la figura anterior se muestra como la tendencia laboral para el siguiente mes es positiva, es decir el modelo pronostica que hay una tendencia del crecimiento del empleo, esto es consistente con la situación actual del país en el que se ha entrado en un proceso de recuperior después del cierre provocado por la pandemia, el valor predictivo es de $27,302,526$ que es un nivel menor al del numero de trabajadores actuales que es de $27,343,451$ por lo que si bien la tendencia que marca el modelo es al alza este valor predicho podría estar subestimado.

1.1.2. Principal Component Regression, PCR

Siguiendo con la idea de que se tienen mas variables de las que realmente necesitamos ahora ocupamos el método de componentes principales los cuál nos permite realizar un análisis de reducción por componentes principales.
La idea del modelo es construir las variables de componentes principales y después usar estos componentes como predictores de en un modelo de regresión lineal el cual es ajustado usando minimos cuadrados, es decir asumimos que solo un pequeño numero de componentes es suficiente para explicar la mayor parte de la variabilidad del modelo. En orden de decidir cual es el numero de componentes primero corremos el método de pcr usando R y graficamos el MSE usando toda la muestra como se observa en la siguiente figura.

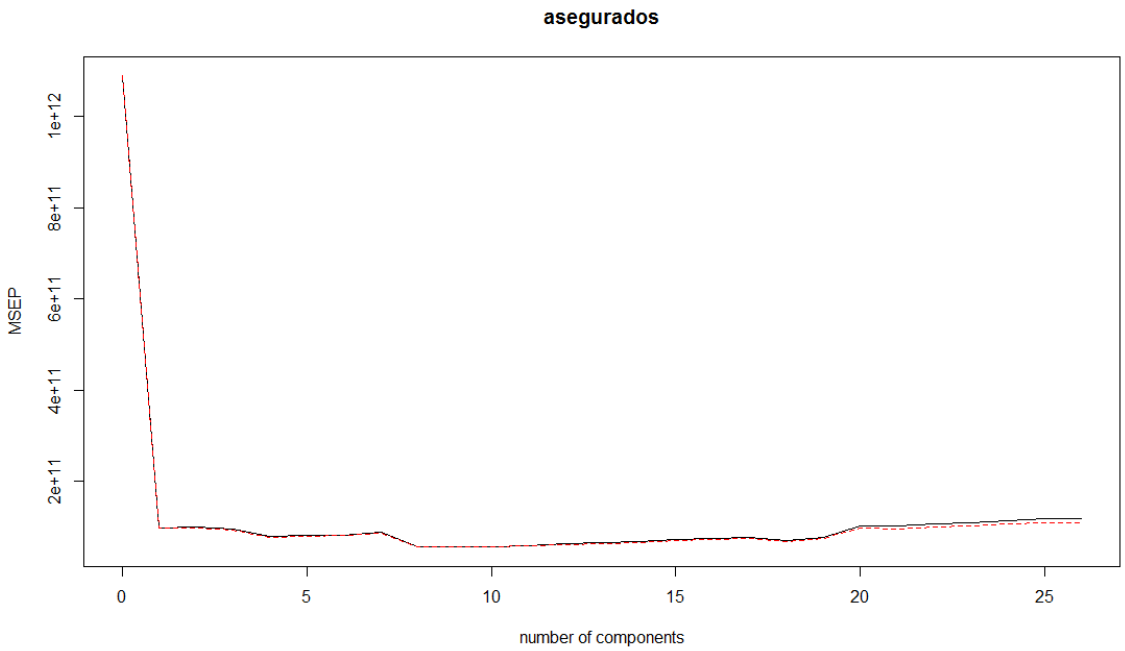

Figura 4: MSE para diferentes numeros de componentes principales

Como se muestra en la figura anterior para obtener el MSE no es necesario tomar todos los componentes de la muestra si no con algunos cuando alrrededor de 7-9, en la siguiente tabla se muestra el proceso de cross-validation con la raíz de MSE y el porcentaje de variación que se obtiene usando diferentes numeros de componentes, como se puede ver el mínimo MSE es usando 8 componentes con el que se obtiene 96.5 % de explicación de la variavilidad de los datos.

Tabla 2: Resumen de Cross-Validation

	Cross Validation				
	6 comp	7 comp	8 comp	9 comp	10 comp
CV	285258	297128	238270	238345	239531
CV adjusted	284021	295384	235891	235897	237170
	TRAINING: % variance explained				
x	92.15	94.96	96.85	97.72	98.39
asegurados	93.92	94.11	96.24	96.3	96.32

Con el fin de poder estimar el test MSE se realizó el mismo procedimiento pero segmentando la muestra en una de entrenamiento y una de test, al igual se muestra la grafica del MSE vs numero de componentes y la tabla de resumen.

Tabla 3: Resumen de Cross-Validation

	Cross Validation				
	6 comp	7 comp	8 comp	9 comp	10 comp
CV	288246	340216	253145	266532	247209
CV adjusted	288411	336150	24817	262272	242506
	TRAINING: % variance explained				
x	94.17	96.09	97.64	98.48	99.03
asegurados	94.4	95.3	96.97	6.97	97.39

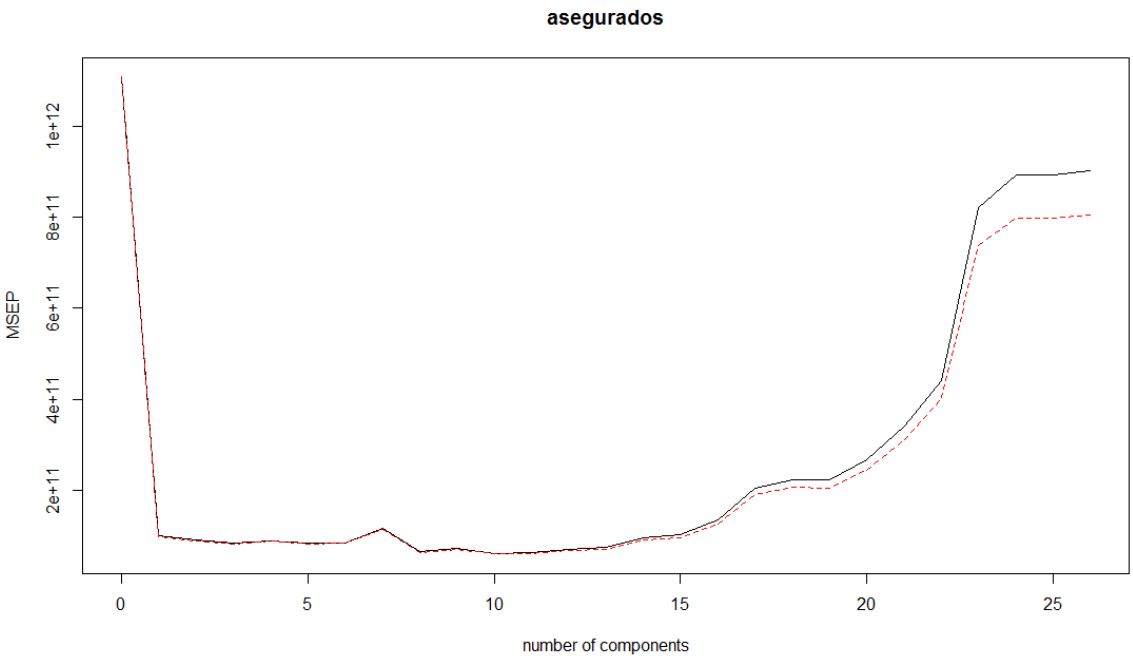

Figura 5: MSE para diferentes numeros de componentes principales

Al igual que usando la muestra completa se comprueba que el mejor modelo se obtiene usando 8 componentes principales con un porcentaje de explicación de 96.97 %.
Realizando la regresión para 8 componentes y realizando la predicción para los regresores regresores de todos los periodos se obtiene el modelo ajustado más la predicción del siguiente mes como se muestra en la siguiente figura.

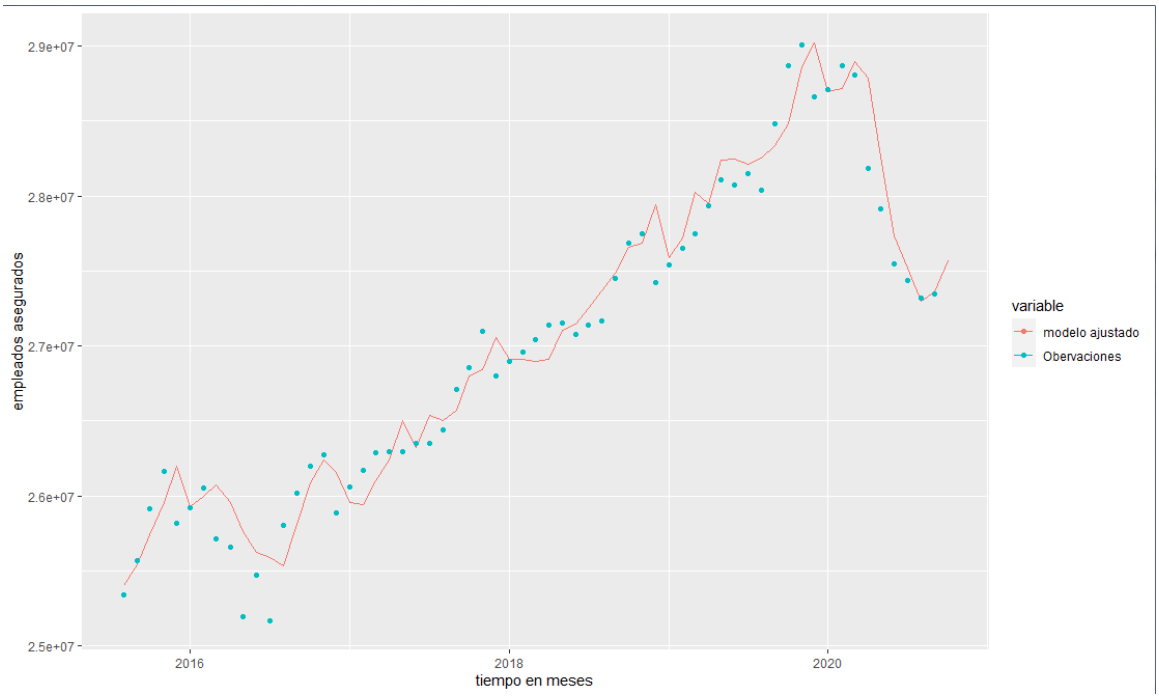

Figura 6: Principal Component Regression

como se observa en la figura anterior el modelo se ajusta bien a lasobservaciones de empleo con las que se cuentan, asimismo se puede ver que la tendencia del siguiente periodo es un incremento en el empleo al igual que el en modelo LASSO sin embargo la pendiente de la curva es mayor cualitativamente en comparación al modelo anterior por lo que bajo este modelo presumiblemente la recuperión del empleo será más rápida, el valor del empleo predicho por el modelo es $27,565,028$ que es mayor al valor actual observado del empleo $27,343,451$ (a diferencia del modelo LASSO que el valor predicho era menor), por lo tanto esta prediccón es presumiblemente mas confiable que la primera, dadas las condiciones

económicas actuales podria pensarse que está predicción esta sobre estimada si tomamos en cuenta que la mayoria de las instituciones del país pronostican una lenta recuperación del empleo y dado que la economía continua parcialmente cerrada se esperaria un incremento más leve, para poder decidir que modelo marca una mejor tendencia analizaremos otro enfoque.

1.1.3. Random Forest

Random Forest es un enfoque donde el numero de arboles de desición se obtiene mediente brootstrap de la muestra de entrenamiento y donde cada arbol es considerado una muestra aleatoria de los predictores. Por lo general se esoge una división de un tercio del numero de predictores que en nuestro caso es alrrededor de 6, entonces corriendo random forest en R se pueden identificar las variables más importantes para explicar la variabilidad del empleo, usando 6 divisiones se puede explicar alrrededor del 93.11 %.

Ahora se presentan los gráficos de importancia donde se puede observar que las variables rezagadas de los asegurados y los trabajadores permanentes urbanos son las variables más importantes (además observe que dos de estas variables coinciden con las encontradas por el método de LASSO por lo que es una forma indirecta de validación entre ambos enfoques), otras variables que también son de importancia son los puestos de trabajo afiliados (como es de esperarse) y la masa salarial.

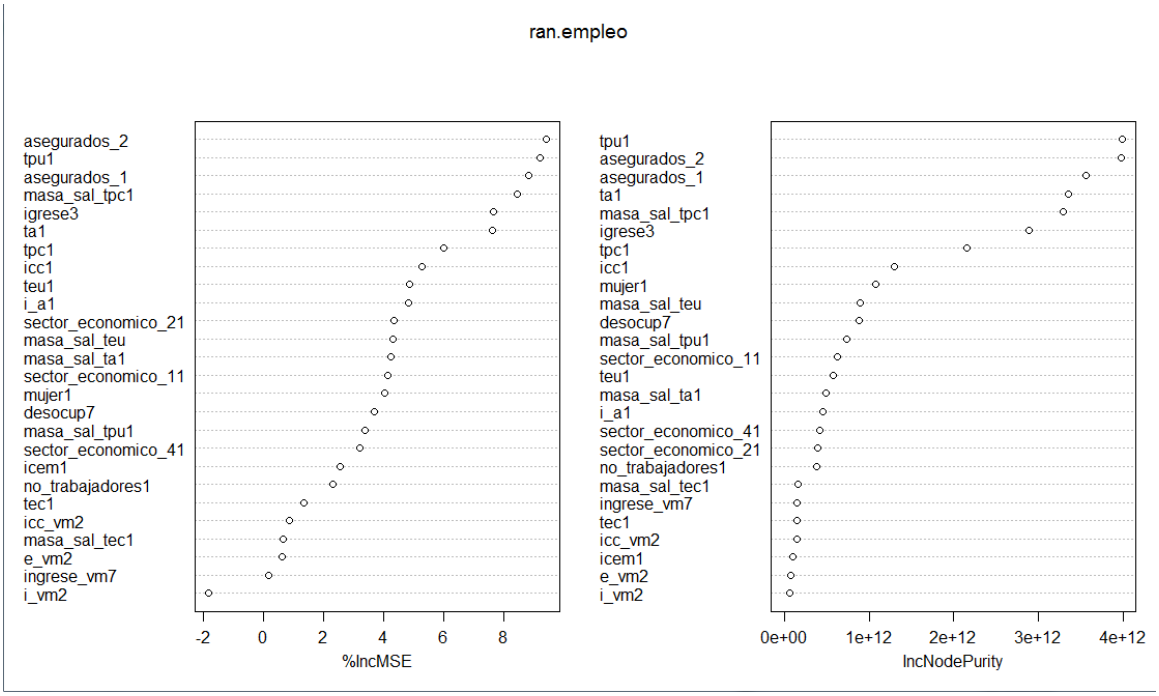

Figura 7: Análisis de importancia

Una vez estimado el modelo con la muestra de entrenamiento se calculó la predicción para la muestra test con el fin de calcular el MSE y realizar la comparación con los otros modelos, sin embargo antes de ellos presentaremos el ajuste para todas las observaciones más la predicción futura para dar una comparación cualitativa con los otros modelos, el grafico de las observaciones y el ajuste se meustra en la siguiente figura.

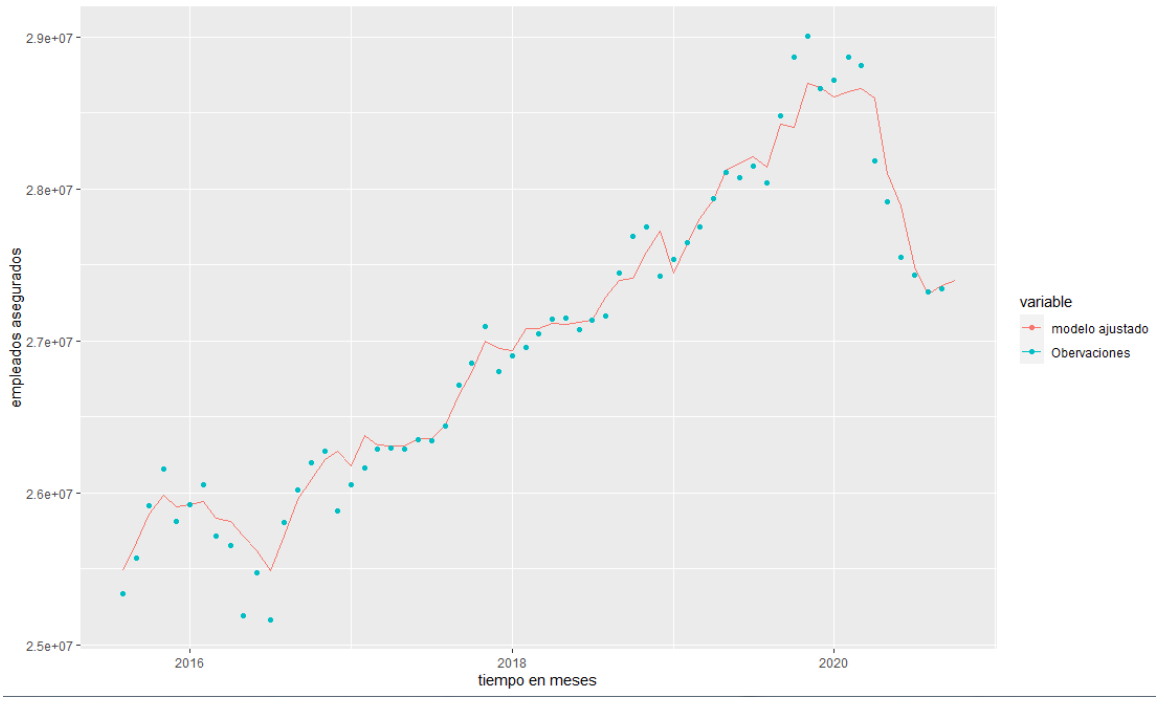

Figura 8: Random Forest

Lo primero que se observa es que la tendencia sobre el empleo al igual que los modelos anteriores es positiva, es decir el modelo estima un crecimiento del empleo para el mes siguiente lo cual es congruente con la situación actual del país, por otro lado cualitativamente este crecimeinto parece ser mucho mas leve que el predicho por los otros modelos (pendiente) lo cual es congruente con los pronósticos realizados por diferentes instituciones sobre una lenta recuperación del nivel del empleo. El nivel de empleo predicho por el modelo es $27,395,715$ comparado con la observación actual que es de $27,343,451$ es un ligero aumento del nivel de empleo (comparado al predicho con PCR).

con el fin de poder elegir el mejor modelo nos basamos en el criterio del menor MSE test estimado, para esto en cada modelo el ajuste se realizó mediante una muestra aleatoria identica y se comparado la predicción con una muestra test (complemento de la muestra de entrenamiento), los MSE test estimados se muestran en la siguiente tabla en donde además se observa que el modelo ganador es Random Forest el cual cualitativamente mediante la figura anterior se observó que era el que mejor modelaba la situación actual del empleo.

Tabla 4: Mejor modelo

	Modelo de Machine Learning		
	LASSO	PCR	Random Forest
MSE test	665,217	653,612	443,209
Predicción	27,302,526	27,565,028	27,395,715
Variables/componentes	4	8	9