EXAMEN PREDICCIÓN

Jorge Casan Vázquez

PREGUNTA 1

A través del test anova obtenemos las variables que son estádisticamente más significativas, las cuales son Valoración_limpieza, Valoración_Tranquilidad, Valoración_Paisajes y Año. Nos quedaremos con nuestro modelo_bueno_training por ser aquel con menor AIC, siendo de un 62.00. Realizando la distancia Cook, nos ayuda a identificar los valores influyentes en nuestro modelo con la desventaja fundamental que no indican comoafectan al modelo.

Con la matriz de confusión se comprobará cómo de bueno es nuestro modelo predictivo. Tenemos que tener en cuenta que estamos trabajando sobre 150 observaciones, tanto para el train como para el test. Además, cada muestra corresponde a un período temporal diferente. Realizamos el cut-off en 0.125 por ser el corte óptimo para poder calcular nuestra matriz de confusión.

Para la valoración del medio ambiente en las Islas Canarias obtenemos una precisión del 76%, en donde la cantidad de falsos positivos son solamente 10 observaciones y la cantidad de falsos negativos de 26 observaciones frente a los verdaderos positivos de 27 observaciones y verdaderos negativos de 87 observaciones, lo cual quiere decir que el porcentaje de la predicción para la valoración del medio ambiente en las Islas Canarias es bajo, siendo el error en un 24%.

Con la Curva ROC representaremos en un gráfico bidimensional la proporción de verdaderos positivos y falsos negativos, dando como resultado un área por debajo de la curva del 0.77.

Continuaremos con nuestro análisis para los modelos de regularización con el objetivo de reducir los coeficientes teniendo como efecto la reducción de la varianza.

Por una parte, a través del modelo Ridge, tienen una penalización por contracción, teniendo como efecto la reducción de las estimaciones del coeficiente hacia cero. Observamos que a media que aumenta el error cuadrático medio aumenta el lambda, sabiendo que cuando se cambia de pendiente es cuando tendremos que elegir el valor. El mínimo error cuadrático medio es de 0.062 y el lambda mínimo de 0.1431. Por otra parte, la selección de variables según este método son Valoración_Limpieza, País_residencia_España y el sexo mujer las que mejor explican la valoración del medio ambiente.

Por otra parte, a través del modelo Lasso, el cual supera la desventaja de Ridge, el cual no incluye todos los predictores en el modelo final ni los fuerza a que sean exactamente cero, tiende generar modelos mucho más interpretables. El error cuadrático medio para este modelo es de 0.05810 y el lambda mínimo 0.04173085. Por otra parte, la selección de variables según este método son Valoración_limpieza y Paisajes.

La elastic net es una combinación de los dos modelos anteriores incorporando la selección de variables del modelo Lasso y la contracción de los predictores correlacionados como en la regresión Ridge.

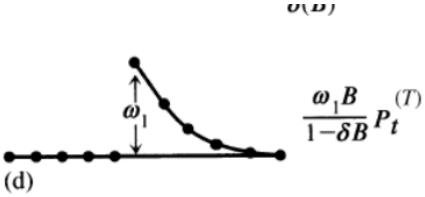
No hay un algoritmo dominante presente aquí, en general es mejor probar las tres técnicas introducida usando estimaciones de error prueba con validación cruzada.

Tomado todo ello en su conjunto, considero que el mejor modelo predictivo es la GLM de la regresión logística por ser la que mejor precisión arroja a nuestro modelo, con una buena curva ROC explicando un porcentaje elevado de verdaderos positivos, siendo las variables que mejor explican la valoración del medio ambiente, las siguientes:

- ✓ Valoración Limpieza
- ✓ Valoración Paisaje
- ✓ Valoración Tranquilidad
- ✓ Año

PREGUNTA 2

Representamos las series temporales según los datos proporcionados en semanas y en meses. Observamos que ambas se comportan como impulso y no escalón puesto que la tendencia de ambas series se recupera en cuando a su media y varianza. Existen muchos picos para las semanas y sin embargo, para la de los meses, la tendencia es más suavizada pese a que tenga haya tenido un crecimiento de las ventas muy pronunciado antes del año 2010. En mi opinión, ambas series siguen el siguiente patrón polinómico, como impulso:



Continuando nuestro análisis vamos a realizar el modelo ARIMA en donde para la semanal tenemos un ARIMA(0,1,0) y para la mensual ARIMA(2,0,2) (2,0,0). Realizando un análisis de los residuos vemos que se comportan como ruido blanco, tanto gráficamente como en el box.test. No obstante, para la serie semanal no arroja resultados concluyentes, solamente la mensual.

Realizando la previsión mensual la predicción para el mes de Agosto es optimista, alcanzando una previsión en ventas de 1653 dólares. Por otra parte, y sabiendo que los resultados no son concluyentes para la serie semanal se espera que las ventas sean iguales para las cuatro semanas de Agosto, siendo de 950 dólares para todas ellas. Para el análisis de los outliers solamente detectamos outliers innovativos para la serie semanal.

Por otra parte, para el modelo ETS en la serie semanal tenemos error aditivo, pero sin tendencia ni componente estacional; y para la serie mensual tenemos error multiplicativo. Tanto en la predicción del modelo ETS para las semanas y para los meses los resultados

son muchos más optimistas que realizando el modelo ARIMA. La predicción para la serie semanal de agosto es de 1696 dólares en ventas y para la mensual de 1859 dólares.

Tomado todo ello en su conjunto, las predicciones tanto semanales como mensuales son mucho más optimistas para el modelo ETS. Por otra parte, al ser la serie estacionaria elegiré realizar la predicción con el modelo ETS ya que ARIMA lo que realiza es convertir una serie no estacionaria en estacionaria a través de la estacionalidad de la media móvil con diferencias y la estacionalidad de la varianza a través de su conversión logarítmica.