REGRESSION FOREST PARA CORRUPCIÓN AMPLIA

Detalles sobre el presente reporte

- Fecha: 21 de febrero de 2024
- Método: Regression Forest
- Nivel de observación: año inicial del reporte por municipalidad y año.
- Variables predictoras: SIAF, Renamu y variables políticas
- Variable predicha: corrupción amplia
- Periodo en el que fue entrenado el modelo: 2016-2020
- Tipo de predicción: clasificación

Resultados (métricas de desempeño)

La tabla 1 presenta los resultados de los modelos de Machine Learning para el conjunto de entrenamiento NRO. Tomando en cuenta la métrica F1, el modelo con el mejor desempeño es el modelo **Regression Forest** entrenado con el conjunto de entrenamiento SMOTE Tomek-Links a un threshold de 0.55.

Tabla 1. Métricas de desempeño de los modelos entrenados con el conjunto de entrenamiento SMOTE Tomek-Links

Métrica	Regresión Logística	Lasso	Ridge	Elastic Net	Random Forest	Gradient Boosting Trees	LGBM Classifier	Regression Forest (threshold = 0.55)
F1	0.486	0.407	0.407	0.407	0.565	0.563	0.525	0.666
Accuracy	0.677	0.494	0.494	0.494	0.911	0.908	0.911	0.872
AUC ROC	0.554	0.630	0.630	0.630	0.757	0.727	0.699	0.770
F1 (Sí)	0.799	0.634	0.634	0.634	0.953	0.952	0.953	0.928
F1 (No)	0.173	0.180	0.180	0.180	0.178	0.174	0.098	0.404
Fuente: elaboración propia								

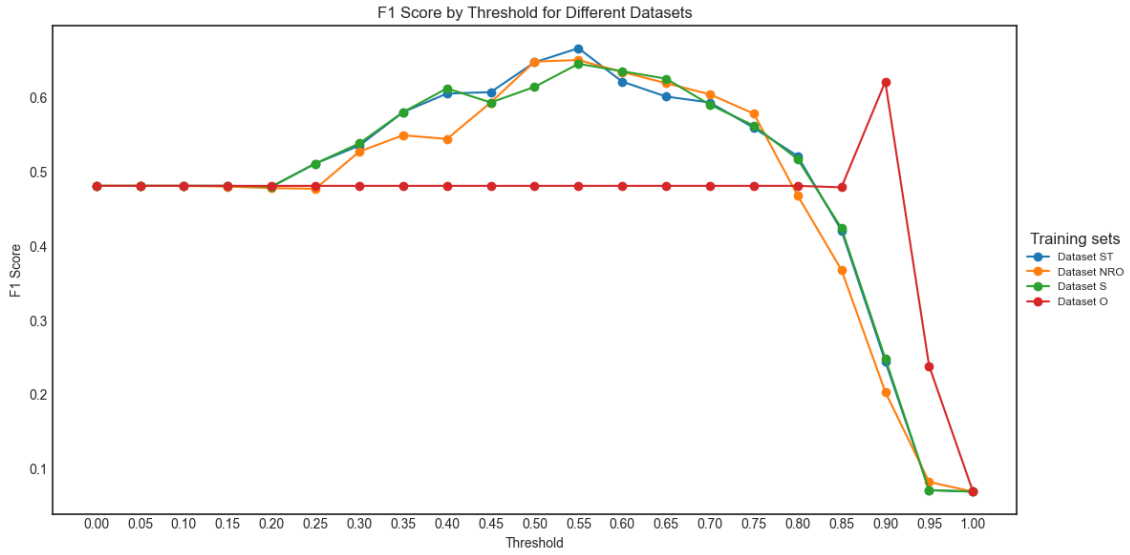
Asimismo, se presentan los 10 modelos con mejor desempeño considerando los distintos métodos de remuestreo empleados (SMOTE, SMOTE Tomek-Links y Naive Random Oversampling). Este ranking se realiza teniendo en cuenta la métrica F1:

Tabla 2. 10 modelos con mayor poder predictivo considerando los distintos métodos de remuestreo

Modelo	F1
Regression Forest ST (threshold = 0.55)	0.666
Regression Forest NRO (threshold = 0.55)	0.650
Regression Forest NRO (threshold = 0.50	0.648
Regression Forest ST (threshold = 0.50)	0.647
Regression Forest S (threshold = 0.55)	0.645
Regression Forest S (threshold = 0.60)	0.635
Regression Forest NRO (threshold = 0.60)	0.634
Regression Forest S (threshold = 0.65)	0.625
Regression Forest O (threshold = 0.90)	0.621
Regression Forest ST (threshold = 0.60)	0.621
Fuente: elaboración propia	

A continuación se presenta un gráfico de cómo varía la métrica F1 en función de distintos thresholds en el rango [0, 1] con pasos de 0.5.

Gráfico 1. F1 en función de distintos thesholds para distintos conjuntos de entrenamiento.



REGRESSION FOREST PARA CORRUPCIÓN INTENSA

Detalles sobre el presente reporte

- Fecha: 21 de febrero de 2024
- Método: Regression Forest
- Nivel de observación: año inicial del reporte por municipalidad y año.
- Variables predictoras: SIAF, Renamu y variables políticas
- Variable predicha: corrupción intensa
- Periodo en el que fue entrenado el modelo: 2016-2020
- Tipo de predicción: clasificación

Resultados (métricas de desempeño)

La tabla 3 presenta los resultados de los modelos de Machine Learning para el conjunto de entrenamiento NRO. Tomando en cuenta la métrica F1, el modelo con el mejor desempeño es el modelo **Regression Forest** entrenado con el conjunto de entrenamiento Naive Random Oversampling a un threshold de 0.4.

Tabla 3. Métricas de desempeño de los modelos entrenados con el conjunto de entrenamiento SMOTE Tomek-Links

Métrica	Regresión Logística	Lasso	Ridge	Elastic Net	Random Forest	Gradient Boosting Trees	LGBM Classifier	Regression Forest (threshold = 0.4)
F1	0.286	0.424	0.424	0.424	0.587	0.567	0.572	0.631
Accuracy	0.316	0.424	0.424	0.424	0.699	0.696	0.687	0.716
AUC ROC	0.524	0.523	0.523	0.523	0.666	0.630	0.614	0.685
F1 (Sí)	0.139	0.440	0.440	0.440	0.802	0.804	0.794	0.808
F1 (No)	0.432	0.407	0.407	0.407	0.372	0.330	0.350	0.454
Fuente: elaboración propia								

Asimismo, se presentan los 10 modelos con mejor desempeño considerando los distintos métodos de remuestreo empleados (SMOTE, SMOTE Tomek-Links y Naive Random Oversampling). Este ranking se realiza teniendo en cuenta la métrica F1:

Tabla 4. 10 modelos con mayor poder predictivo considerando los distintos métodos de remuestreo

Modelo	F1
Regression Forest NRO (threshold = 0.40)	0.631
Random Forest Classifier SMOTE	0.627
Regression Forest O (threshold = 0.70)	0.626
Regression Forest NRO (threshold = 0.45)	0.622
Regression Forest ST (threshold = 0.50)	0.617
Regression Forest NRO (threshold = 0.55)	0.616
Regression Forest S (threshold = 0.45)	0.613
Regression Forest O (threshold = 0.65)	0.611
Regression Forest S (threshold = 0.55)	0.611
Random Forest Classifier ST	0.609
Fuente: elaboración propia	

A continuación se presenta un gráfico de cómo varía la métrica F1 en función de distintos thresholds en el rango [0, 1] con pasos de 0.5.

Gráfico 2. F1 en función de distintos thesholds para distintos conjuntos de entrenamiento.