# Proyecto II:
# Modelo predictivo en R.
# Washington Loans

Paula Durá
Francisco Jara
Manuel Marín

Universitat d'Alacant
Universidad de Alicante

# Resumen ejecutivo

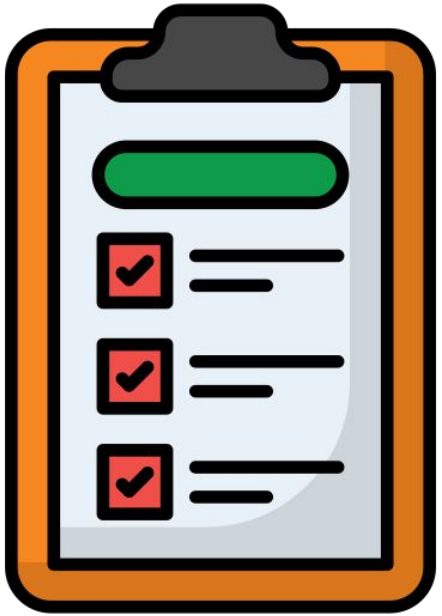
Como consultora financiera, nuestro **objetivo principal** es ayudar a la entidad prestamista en el proceso de **selección de préstamos hipotecarios**, buscando siempre la disminución de los riesgos y el aumento de la rentabilidad, de manera que se cumpla la normativa impuesta a la empresa y se estudien las nuevas tendencias y necesidades del mercado.

Para ello, se han analizado a través de R todas las hipotecas solicitadas en el estado de Washington (EE.UU) a lo largo del año 2016 y se han llegado a las siguientes conclusiones:

- Descartamos dos modelos iniciales, uno con todas las variables y otro seleccionando las más significativas del modelo anterior, debido a su alta dimensionalidad y, por tanto, su difícil interpretación. Después, descartamos un tercer modelo, obtenido a partir de las variables con mejores p-valores de regresiones logísticas individuales, y por su alta tasa de falsos positivos.
- Seleccionamos un modelo de manera manual, en base a los conocimientos del campo, correlaciones y análisis exploratorios previos. En él detectamos el 66,5% de los verdaderos positivos y se seleccionan 7 variables: tipo de propiedad, minorías raciales, ingresos anuales del solicitante, monto del préstamo, propósito del préstamo, gravamen y tipo de préstamo.
- Una vez tenemos nuestro modelo final, interpretamos los resultados que nos proporciona R. Los préstamos solicitados para casas prefabricadas, refinanciación, sin gravamen y con segundo gravamen disminuyen la probabilidad de aprobación un 33’23%, 33’9%, 37% y 21%, respectivamente, Además, si el solicitante pertenece a una zona de minoría racial, la probabilidad de que el préstamo sea aceptado disminuye un 0.42%. En cambio, si el préstamo es para una compra de vivienda, las probabilidades de aprobación aumentan un 70%.
- En vista de ello, se recomienda a la entidad financiera que busque un perfil adecuado basado en estos resultados: con altos ingresos, que no habite en zonas de minoría racial y que solicite un préstamo para la compra de una vivienda unifamiliar o multifamiliar con primer gravamen.

# Agenda

- Introducción
- Modelos descartados
- Selección del modelo final
- Interpretación del modelo final
- Predicciones y recomendaciones
- Next Steps
- Backup

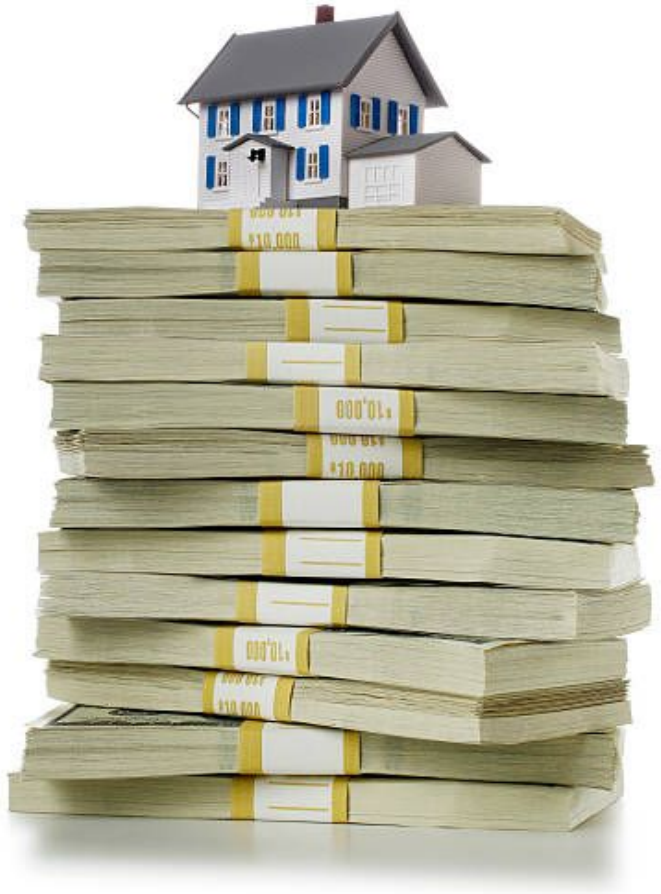


# Introducción

En el sector financiero actual, la toma de decisiones crediticias requiere un equilibrio entre riesgo y rentabilidad. Este estudio analiza 466,000 solicitudes del dataset HMDA de Washington State mediante técnicas estadísticas en R.

El objetivo principal del proyecto es predecir si se va a conceder o no un préstamo, identificando relaciones en la aprobación de préstamos y evaluar el impacto de variables clave; como ingresos, tipo de propiedad y perfil demográfico.

Este estudio no solo revela patrones ocultos en los datos, sino que abre la puerta a oportunidades de mercado: al optimizar los criterios de aprobación, las instituciones pueden ampliar su cartera de préstamos de manera segura, impulsando el acceso a vivienda y estimulando el crecimiento económico local.



# Modelos descartados

| Modelo utilizando todas las variables | Modelo con las variables significativas del primer modelo | Modelo tomando el top 10 de los p-valores de regresiones logísticas individuales |
|---|---|---|
| • Modelo en el que enfrentamos la variable objetivo (aprobación) contra las 21 variables útiles de nuestro dataset.<br>• Este modelo no es bueno por su difícil interpretación debido a la gran cantidad de variables. | • A partir del modelo 1, seleccionamos las variables que nos salen con un p-valor menor de 0’05.<br>• Aunque hayamos conseguido disminuir la dimensionalidad seleccionando las variables significativas, el modelo sigue siendo muy difícil de interpretar porque todavía hay un gran número de variables. | • Enfrentamos la variable objetivo contra las 21 variables explicativas con regresiones logísticas individuales<br>• Tras hacer las regresiones logísticas, tomamos el top 10 p-valores y generamos el modelo<br>• Este modelo reduce la dimensionalidad bastante, pero nos da una tasa de FP de 0’685 . |

# Selección del modelo final

**VARIABLES SELECCIONADAS**

1. **Tipo de propiedad:** casas prefabricadas, casas unifamiliares, edificios
2. **Minorías raciales**
3. **Ingresos anuales del solicitante**
4. **Monto del préstamo**
5. **Propósito del préstamo:** refinanciación, compra de vivienda, reforma
6. **Gravamen:** No tiene, primer gravamen, segundo gravamen
7. **Tipo de préstamo:** FHA, FSA, VA

- Seleccionamos las variables basándonos en las correlaciones con la variable objetivo y el EDA realizado en el trabajo anterior.
- Casi todas las variables empleadas nos salen significativas, por lo que son de gran importancia para la aprobación del préstamo.
- Conseguimos disminuir con el modelo actual la tasa de falsos positivos a un 33’5%, es decir, **detectamos el 66’5% de los verdaderos positivos**.

# Interpretación del modelo

| Variable | Interpretación | Propensión a ser aprobado |
|---|---|---|
| Tipo de propiedad: casa prefabricada | Las viviendas prefabricadas tienen 34.23% menos probabilidad de aprobación que las viviendas unifamiliares o los bloques de viviendas. | 34.23% ↓ |
| Minorías | Cada aumento del 1% en la población minoritaria, reduce la probabilidad de aprobación un 0.42%, lo que se traduce que un área con 50% de población minoritaria, tiene 21.5% menos de probabilidad de aprobación. | 0.42% ↓ |
| Propósito: compra de vivienda y refinanciación | Los préstamos para compra de la vivienda aumentan un 75% la probabilidad de ser aprobados con respecto a aquellos con fines de reforma. Por otro lado, los destinados a refinanciaciones disminuyen la probabilidad aprobación un 33.9%. | **Compra de vivienda:** 75% ↑<br>**Refinanciación:** 33.9% ↓ |
| Gravamen | Los préstamos sin gravamen y de segundo gravamen disminuyen la probabilidad de ser aprobados con respecto a los que tienen un primer gravamen | **Sin gravamen:** 37% ↓<br>**Segundo gravamen:** 21% ↓ |

# Conclusiones y recomendaciones para la entidad financiera

## PERFIL IDEAL

- ➔ Vive en áreas no minoritarias
- ➔ Opta por viviendas unifamiliares o multifamiliares
- ➔ El propósito de su préstamo es la compra de una vivienda
- ➔ Solicitan préstamos con primer gravamen
- ➔ Tienen mayores ingresos anuales

## PERFIL CRÍTICO

- ➔ Vive en áreas minoritarias
- ➔ Opta por casas prefabricadas
- ➔ El propósito de su préstamo es la refinanciación
- ➔ Solicitan préstamos sin gravamen
- ➔ Tienen menores ingresos anuales

# Next steps

- Mejora de predicciones usando métodos más complejos que se adhieran mejor a la estructura de los datos.

- Segmentación de ofertas: personalizar las condiciones del préstamo según el perfil del solicitante

- Revisar o ajustar los criterios de aprobación cuando se identifiquen factores que sistemáticamente disminuyan la probabilidad de aprobación, asegurándose de que no se descarten oportunidades valiosas de negocio

# BACKUP

# Backup I: Modelos descartados

## MODELO 1

```
glm(formula = aprobacion ~ Tract + rate_spread + minorias + nro_viv_ocupadas +
    nro_unid_residenciales + loan_amount_000s + Ingreso_medio_fam +
    Ingreso_solicit + tipo_comprador + tipo_propiedad + preaprovacion +
    tipo_ocupacion + loan_type_name + proposito + gravamen +
    edit_status_name + sexo_coapp + etnia_coapp + Raza + Etnia +
    agency_abbr, family = "binomial", data = variables)
```

## MODELO 2

```
glm(formula = aprobacion ~ Tract + minorias + nro_viv_ocupadas +
    nro_unid_residenciales + loan_amount_000s + Ingreso_solicit +
    tipo_propiedad + preaprovacion + tipo_ocupacion + loan_type_name +
    proposito + gravamen + edit_status_name + sexo_coapp + etnia_coapp +
    Raza + Etnia + agency_abbr, family = binomial, data = variables)
```

## MODELO 3

```
glm(formula = aprobacion ~ loan_amount_000s + Tract + Ingreso_medio_fam +
    Etnia + msamd_name + Ingreso_solicit + etnia_coapp + minorias +
    Sexo + Etnia, family = binomial, data = datos)
```

# Backup II: Información del modelo seleccionado

```
Call:
glm(formula = aprobacion ~ tipo_propiedad + minorias + Ingreso_solicit *
    loan_amount_000s + proposito + gravamen + Ingreso_solicit *
    loan_type_name, family = binomial, data = variables)

Coefficients:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    7.741e-01  1.915e-02  40.410  < 2e-16 ***
tipo_propiedadManufactured housing            -4.149e-01  2.023e-02 -20.512  < 2e-16 ***
minorias                                      -4.281e-03  2.268e-04 -18.872  < 2e-16 ***
Ingreso_solicit                                6.427e-04  4.821e-05  13.332  < 2e-16 ***
loan_amount_000s                               1.495e-04  2.481e-05   6.024 1.70e-09 ***
propositoHome purchase                         5.624e-01  1.772e-02  31.741  < 2e-16 ***
propositoRefinancing                          -4.136e-01  1.730e-02 -23.915  < 2e-16 ***
gravamenNo tiene                              -4.620e-01  3.286e-02 -14.062  < 2e-16 ***
gravamensubord                                -2.368e-01  2.409e-02  -9.830  < 2e-16 ***
loan_type_nameFHA-insured                     -4.224e-01  1.965e-02 -21.496  < 2e-16 ***
loan_type_nameFSA/RHS-guaranteed              -4.677e-01  1.094e-01  -4.275 1.91e-05 ***
loan_type_nameVA-guaranteed                   -2.830e-01  2.246e-02 -12.600  < 2e-16 ***
Ingreso_solicit:loan_amount_000s              -2.488e-07  3.306e-08  -7.525 5.26e-14 ***
Ingreso_solicit:loan_type_nameFHA-insured      3.711e-04  2.082e-04   1.783 0.074646 .
Ingreso_solicit:loan_type_nameFSA/RHS-guaranteed 1.279e-03  1.703e-03   0.751 0.452524
Ingreso_solicit:loan_type_nameVA-guaranteed    7.658e-04  2.169e-04   3.531 0.000414 ***
```

```
    Null deviance: 488987  on 381428  degrees of freedom
Residual deviance: 470305  on 381413  degrees of freedom
  (510 observations deleted due to missingness)
AIC: 470337
```

```
Matriz de Confusión:
          Predicho
Real            0     1
  Denegado  24898 12542
   aprobado 33098 39166

Tasa de falsos positivos (Error Tipo 1): 0.335
Tasa de falsos negativos (Error Tipo 2): 0.458
```

0.0382063262633145

umbral de p-valor: 0.62