Universidad Continental

FACULTAD DE INGENIERÍA

Escuela Académico Profesional de Ingeniería de Sistemas e Informática

Tesis

Modelo basado en patrones para la evaluación de recuperabilidad en cartera atrasada de una entidad bancaria, Arequipa 2020

Manuel Enrique Arenas Cáceres

Para optar el Título Profesional de
Ingeniero de Sistemas e Informática

Arequipa, 2020

ÍNDICE

AGRADECIMIENTOS ii
DEDICATORIA iii
ÍNDICE iv
ÍNDICE DE TABLAS vii
ÍNDICE DE FIGURAS viii
RESUMEN xi
ABSTRACT xii
INTRODUCCIÓN 1
CAPíTULO I 3
PLANTEAMIENTO DEL ESTUDIO 3
1.1. Planteamiento y formulación del problema 3
1.1.1. Planteamiento del problema 3
1.1.2. Formulación del problema 6
1.2. Objetivos 7
1.2.1. Objetivo general 7
1.2.2. Objetivos específicos 7
1.3. Justificación e importancia 7
1.4. Hipótesis y descripción de variables 8
1.4.1. Hipótesis 8
1.4.2. Variables 9
CAPÍTULO II 10
MARCO TEÓRICO 10
2.1. Antecedentes del problema 10
2.1.1. Tesis internacionales 10
2.1.2. Tesis nacionales 11
2.2. Bases teóricas 12
2.2.1. Modelo basado en patrones 12
2.2.1.1. Inteligencia artificial 12
2.2.1.2. Machine Learning 16
2.2.1.3. Random Forest 19
2.2.1.4. Regresión Logística 20

2.2.1.5. Metodología Modelo CRISP-DM 22
2.2.1.6. Dataset 28
2.2.1.7. ETL 31
2.2.1.8. Python 33
2.2.1.9. Power BI 34
2.2.2. Cartera atrasada 36
2.2.2.1. Entidad financiera 36
2.2.2.2. Cartera atrasada 36
2.3. Definición de términos básicos 38
2.3.1. Modelo basado en patrones 38
2.3.2. ETL 38
2.3.3. Random Forest 38
2.3.4. Regresión Logística 38
2.3.5. Accuracy 38
2.3.6. Matriz de confusión 39
2.3.7. Cartera vencida 39
2.3.8. Morosidad 39
2.3.9. Cartera judicial 39
2.3.10. Cartera castigada 39
CAPíTULO III 40
METODOLOGÍA 40
3.1. Método y alcance de la investigación 40
3.1.1. Método de la investigación 40
3.1.2. Alcance de la investigación 40
3.2. Diseño de la investigación 41
3.2.1. Diseño de la investigación 41
3.3. Población y muestra 42
3.3.1. Población 42
3.3.2. Muestra 42
3.4. Técnicas e instrumentos de recolección de datos 43
3.4.1. Técnicas 43
3.4.2. Instrumentos de recolección de datos 43
CAPÍTULO IV 48
IMPLEMENTACIÓN 48

4.1. Aplicación de la Metodología CRISP – DM 48
4.1.1. Primera fase: Business Understanding 49
4.1.2. Segunda fase: Data Understanding 55
4.1.3. Tercera fase: Data Preparation 68
4.1.4. Cuarta fase: Modeling 82
4.1.5. Quinta fase: Evaluation 85
4.1.6. Sexta Fase: Deployment 97
CAPÍTULO V 99
RESULTADOS Y DISCUSIÓN 99
5.1. Resultados del tratamiento 99
5.2. Pruebas de hipótesis 100
5.2.1. Prueba estadística 100
5.2.2. Prueba de hipótesis general 100
5.2.3. Prueba de hipótesis específicas 104
5.3. Discusión de resultados 105
CONCLUSIONES 108
REFERENCIAS BIBLIOGRÁFICAS 109
Anexo 1. Operacionalización de variables 114
Anexo 2. Matriz de consistencia 115
Título: Modelo basado en patrones para la evaluación de recuperabilidad en cartera atrasada de una entidad bancaria, Arequipa 2020 115
Anexo 3. Descripción de costos 116
Anexo 4. Cronograma de actividades de la metodología CRISP-DM. 117
Anexo 5. Descripción de tablas 118
Anexo 6. Código SQL para la creación y Exploración de datos 123
Anexo 7. Código Python Modelos Basados en Patrones 130

ÍNDICE DE TABLAS

Tabla 1 Descripción de la Población por tipo de deuda elaboración propia 42

Tabla 2 Tabla de puntuación Likert para el modelo basado en patrones. 46

Tabla 3 Tabla de % de recuperación por tipo de cartera atrasada Fuente (29) 47

Tabla 4 puntuación Likert para la evaluación de recuperabilidad en cartera atrasada Castigada 47

Tabla 5 Tabla Resumida de las etapas y tiempo de ejecución Elaboración Propia 53

Tabla 6 Datos para la tabla de Dimensión de Asignación..... 70

Tabla 7 Datos para la tabla de Dimensión de Gestión..... 70

Tabla 8 Datos para la tabla de Dimensión de Cuota..... 69

Tabla 9 Datos para la tabla de Dimensión de Pago..... 70

Tabla 10 Tabla de configuración para el algoritmo Regresión Logística.. 84

Tabla 11 Tabla de configuración para el algoritmo Random Forest..... 84

Tabla 12 Comparación de modelos Regresión Logística..... 94

Tabla 13 Comparación de modelos Random Forest..... 94

Tabla 14 Comparación de Modelo Regresión Logística y Random Forest 95

Tabla 15 Comparación de Modelo de Datos para Cada Algoritmo 96

Tabla 16 Estadísticos descriptivos. Porcentaje de Recupero 101

Tabla 17 Coeficiente de correlación intraclase del MODELO 1 102

Tabla 18 Coeficiente de correlación intraclase del MODELO 3 103

Tabla 19 Estructura de la Tabla Asignación 118

Tabla 20 Estructura de la Tabla Pagos..... 120

Tabla 21 Estructura de la Tabla Gestiones..... 120

Tabla 22 Estructura de la Tabla Cuotas..... 122

ÍNDICE DE FIGURAS

Figura 1 Crisis Economicas Mundiales y años en solucionar la morosidad causada por la crisis Fuente (1)..... 3

Figura 2 Proyecciones de crecimientos económico disminuyen y la dispersión de crecimiento aumenta Fuente (2)..... 4

Figura 3 Indice de morosidad en los ceditos en porcentaje Fuente (3)..... 5

Figura 4 Naturaleza Multidisciplinaria de Minería basada en Datos, Inteligencia Artificia y Ciencia de Datos Fuente (12)..... 17

Figura 5 Random Forest formado por cuatro árboles de decisión. Una instancia (línea superpuesta) se clasifica como ideal Fuente (7)..... 20

Figura 6 Modelo de Regresión Logística versus Modelo Regresión Lineal Fuente (16) 22

Figura 7 Separación en cuatro niveles de la metodología CRISP-DM Fuente (18) 23

Figura 8 Las 6 Fases del modelo de proceso CRISP-DM Fuente (19)..... 24

Figura 9 Resumen de tipos de datos Fuente (20)..... 29

Figura 10 Esquema Descriptivo - Correlacional..... 41

Figura 11 Imagen de una Matriz de Confusión 44

Figura 12 Formula para calcular la Tasa de Error y Tasa de acierto 45

Figura 13 Formula para calcular la Precisión..... 45

Figura 14 Formula para Calcular el Porcentaje de Recuperabilidad..... 46

Figura 15 Tablas principales para la investigación 56

Figura 16 Gráfico del número de cuentas asignadas por mes de asignación 57

Figura 17 Gráfico del número de Gestiones Contacto Efectivo por mes de asignación..... 58

Figura 18 Gráfico de número de cuentas por tipo de producto agrupado 59

Figura 19 Gráfico de número de cuentas por Zona 60

Figura 20 Gráfico de número de cuentas por departamento 60

Figura 21 Gráfico de número de cuentas por tipo de Moneda..... 61

Figura 22 Gráfico de número de cuentas por Porcentaje de descuento. . 62

Figura 23 Gráfico de número de cuentas por mes de asignación y tipo de segmento 63

Figura 24 Gráfico de número de cuentas con calificación en la SBS..... 63

Figura 25 Gráfico de número de cuentas por mes de asignación y grupo de campaña. 64

Figura 26 Gráfico de número de cuentas por mes de asignación retiro de AFP. 65
Figura 27 Gráfico de cuotas agendadas por mes de asignación. 65
Figura 28 Gráfico de número de acuerdos por estado de acuerdo 66
Figura 29 Gráfico de Cantidad de pagos por mes asignación 67
Figura 30 Creación de tabla dimensión de pagos 72
Figura 31 Creación de tabla dimensión de gestiones 72
Figura 32 Creación de tabla dimensión de cuotas 73
Figura 33 Creación de tabla dimensión asignación 74
Figura 34 Query para Creación de tabla Data Deudas 75
Figura 35 Estructura de los ETL 76
Figura 36 ETL para el modelo de datos 1 en Visual Studio 2019 77
Figura 37 Query del ETL del modelo de datos 1 77
Figura 38 ETL para el modelo de datos 2 en Visual Studio 2019 78
Figura 39 Query del ETL del modelo de datos 2 78
Figura 40 ETL para el modelo de datos 3 en Visual Studio 2019 79
Figura 41 Query del ETL del modelo de datos 3 79
Figura 42 Código en Python para modificar datos de Zona, Grupo Campaña y Prioridad de Gestión 80
Figura 43 Código en Python para modificar datos producto 81
Figura 44 Script en Python para modificar datos producto 81
Figura 45 Lista de librerías importadas de Python 83
Figura 46 Resultados del Modelo de datos 1 Regresión Logística 86
Figura 47 Matriz de confusión del modelo de datos 1 Regresión Logística 87
Figura 48 Resultados del Modelo de datos 2 Regresión Logística 87
Figura 49 Matriz de confusión del modelo de datos 2 Regresión Logística 88
Figura 50 Resultados del Modelo de datos 3 Regresión Logística 88
Figura 51 Matriz de confusión del modelo de datos 3 Regresión Logística 89
Figura 52 Resultados del Modelo de datos 1 Random Forest 90
Figura 53 Matriz de confusión del modelo de datos 1 Random Forest 90
Figura 54 Resultados del Modelo de datos 2 Random Forest 91
Figura 55 Matriz de confusión del modelo de datos 2 Random Forest 92

Figura 56 Resultados del Modelo de datos 3 Random Forest 92
Figura 57 Matriz de confusión del modelo de datos 3 Random Forest 93
Figura 58 Gráfica de datos del Modelo 1 102
Figura 59 Gráfica de datos del Modelo 2 104

RESUMEN

El presente trabajo de tesis tiene como objetivo determinar la relación del modelo basado en patrones con la evaluación de recuperabilidad en cartera atrasada de una entidad bancaria. Se plantearon los objetivos específicos: medir la relación del modelo basado en Regresión Logística para la evaluación de recuperabilidad en cartera atrasada y medir la relación del modelo basado en Random Forest para la evaluación de recuperabilidad en cartera atrasada. Las predicciones de pago del modelo y los pagos reales efectuados en la cartera atrasada tuvieron un coeficiente de correlación intraclase, con nivel de insignificancia 0.00 que evidenció un alto nivel de relación. El modelo basado en patrones pronosticó un 16 % de recuperabilidad que se considera bueno, según la escala de puntuación Likert de recuperabilidad de cartera atrasada. El algoritmo Random Forest obtuvo un Accuracy de 0.92 y una Precisión de 0.71, para el algoritmo de Regresión Logística se obtuvo un Accuracy de 0.90 y una Precisión de 0.44, ambos tuvieron un Accuracy mayor a 0.90 considerados muy buenos en la escala de puntuación Likert del modelo basado en patrones. Cumpliendo todos los objetivos se afirma una fuerte relación entre el modelo basado en patrones y la evaluación de recuperabilidad de cartera atrasada de una entidad bancaria.

Palabras claves: modelo basado en patrones, cartera atrasada, entidad financiera, algoritmos Random Forest y Regresión Logística, Accuracy.

ABSTRACT

The objective of the present thesis work is to determine the relationship of the pattern-based model with the assessment of recoverability in the delinquent portfolio of a bank. The specific objectives were raised; measure the relationship of the model based on Logistic Regression for the evaluation of recoverability in arrears and measure the relationship of the model based on Random Forest for the evaluation of recoverability in arrears. The payment predictions of the Model and the actual payments made in the delinquent portfolio had an intraclass correlation coefficient, with an insignificance level of 0.00, which evidenced a high level of relationship. The pattern-based model predicted a 16% recoverability that is considered good according to the Likert scoring scale for recoverability of arrears. The Random Forest algorithm obtained an Accuracy of 0.92 and a Precision of 0.71, for the Logistic Regression algorithm an Accuracy of 0.90 and a Precision of 0.44 were obtained, both had an Accuracy greater than 0.90 considered very good in the Likert scoring scale of the pattern-based model. Fulfilling all the objectives, a strong relationship is affirmed between the pattern-based model and the evaluation of the recoverability of the delinquent portfolio of a banking entity.

Keywords: Pattern-Based Model, Overdue Portfolio, Financial Institution, Random Forest Algorithms and Logistic Regression, Accuracy.