Souphanouvong University Journal of Multidisciplinary Research and Development
2025, Vol. 11, No. 03, 01-10
https://doi.org/10.69692/SUJMRD110301

Comparative Analysis of Logistic Regression and Random Forest Models for Customer Churn Prediction in Laos

Sithong PHETVILAI[*1], Savath SAYPADITH[2], Vimontha KHIEOVONGPHACHANH[3]
Computer Engineering and Information Technology Department, Faculty of Engineering, National University of Laos, Lao PDR

***Correspondence:** Sithong phetvilai, Computer Engineering and Information Technology Department, Faculty of Engineering, National University of Lao, +8562078885566, sithongphetvilai@gmail.com

Article Info:
Submitted: **June 04, *2025***
Revised: June 29, 2025
Accepted: July 09, 2025

Abstract

Customer churn poses a significant challenge in the telecommunications industry, as it directly impacts both revenue and long-term customer retention. This study leverages real-world customer data from TPLUS Digital to evaluate the effectiveness of machine learning techniques in predicting customer churn. The research aims to assess the learning speed and testing capability of machine learning models and compare their performance in churn prediction. Two widely used models Logistic Regression (LR) and Random Forest (RF) were employed and evaluated using various performance metrics, including training and testing time, AUC, Precision, Recall, and F1-Score. The dataset was divided into training and testing sets using multiple ratios: 70%-30%, 80%-20%, and 90%-10%. Results show that Random Forest outperforms Logistic Regression in identifying customers likely to churn and achieves higher overall accuracy. However, Logistic Regression exhibited faster training times and more consistent performance on imbalanced datasets. This study provides valuable insights into the application of machine learning for churn prediction in the telecommunications sector and offers a foundation for developing effective customer retention strategies.

Keywords: Logistic Regression, Random Forest, Churn Prediction, Model Evaluation Metrics

1. ພາກສະໜີ

ໃນຍຸກປັດຈຸບັນນີ້ ເປັນຍຸກທີ່ເຕັກໂນໂລຊີ ແລະ ຂໍ້ມູນຂ່າວສານຖືກນຳມາໃຊ້ສະໜັບສະໜຸນສັງຄົມໃນຫຼາຍດ້ານ. ໂດຍສະເພາະ ການເຂົ້າເຖິງ ແລະ ຄົ້ນຫາຂໍ້ມູນຂ່າວສານຜ່ານລະບົບ ໂທລະຄົມມະນາຄົມ ຈຶ່ງກ່າຍມາເປັນເລື່ອງງ່າຍ, ສະດວກ ແລະ ວ່ອງໄວຂຶ້ນເລື້ອຍໆ. ລະບົບໂທລະຄົມມະນາຄົມໄດ້ມີການພັດທະນາ ແລະ ຂະຫຍາຍຕົວຢ່າງກ້າວກະໂດດ ເຊິ່ງອາດຈະເປັນລະບົບທີ່ມີສາຍ, ບໍ່ມີສາຍ ຫຼື ເຄືອຂ່າຍ ອິນເຕີເນັດ ເປັນຕົ້ນ. ສິ່ງເຫຼົ່ານີ້ຊ່ວຍໃຫ້ຜູ້ໃຊ້ບໍລິການສາມາດສົ່ງ ແລະ ຮັບຂໍ້ມູນຂ່າວສານໄດ້ຢ່າງວ່ອງໄວ ທັນເວລາ ແລະ ຄວບຄຸມທຸກຮູບແບບການສື່ສານ ບໍ່ວ່າຈະເປັນສຽງ, ຂໍ້ມູນທີ່ເປັນຕົວເລກ, ຕົວອັກສອນ, ຮູບພາບ, ວິດີໂອ ຫຼື ແຜນວາດຕ່າງໆ. ເພາະສາມາດເຜີຍແຜ່ຂໍ້ມູນຂ່າວສານເຫຼົ່ານີ້ໄປຫາຜູ້ໃຊ້ງານຕ່າງໆໄດ້ຢ່າງບໍ່ຈຳກັດ ບໍ່ວ່າຈະຢູ່ໃກ້ ຫຼື ຢູ່ໄກ ເນື່ອງຈາກການພັດທະນາຂອງໂທລະຄົມມະນາຄົມ ແລະ ການນຳອິນເຕີເນັດທີ່ໄດ້ມາຊ່ວຍໃນການໂອນຖ່າຍ ແລະ ສື່ສານໂດຍຜ່ານສັນຍານຕ່າງໆ ທີ່ມີສາຍ ຫຼື ບໍ່ມີສາຍ ຢູ່ໃນອັດຕາຄວາມໄວທີ່ສູງ (Kingsakda et al., 2024).

ແນ່ນອນໃນອຸດສະຫະກຳໃດທີ່ມີການນຳໃຊ້ຫຼາຍກໍ່ເປັນທີ່ຍອມຮັບກັນແລ້ວວ່າ ຕ້ອງມີການແຂ່ງຂັນສູງ ໂດຍຜູ້ໃຫ້ບໍລິການເພື່ອໃຫ້ລູກຄ້າເລືອກໃຊ້ງານ ຄວາມທ້າທາຍທີ່ສຳຄັນອີກຢ່າງໜຶ່ງເຖິງການຮັກສາລູກຄ້າເກົ່າທີ່ມີຢູ່ແລ້ວ ບໍ່ໃຫ້ຍ້າຍໄປຫາຄູ່ແຂ່ງອື່ນ. ສຳລັບອຸດສາຫະກຳການສື່ສານ ການເສຍລູກຄ້າໃຫ້ກັບບໍລິສັດຄູ່ແຂ່ງທີ່ມີຂໍ້ສະເໜີດີກວ່າ ເຮັດໃຫ້ເສຍໂອກາດສຳຄັນທາງທຸລະກິດ ແລະ ກ່ຽວຂ້ອງກັບການສູນເສຍລາຍຮັບ ເນື່ອງຈາກຄ່າໃຊ້ຈ່າຍໃນການຊອກຫາລູກຄ້າໃໝ່ຈະມີຄ່າໃຊ້ທີ່ຈ່າຍສູງກວ່າການຮັກສາລູກຄ້າເກົ່າ (Wagh & Andhale, 2024).

ການຄາດຄະເນການຢຸດໃຊ້ບໍລິການຂອງລູກຄ້າແມ່ນ ມີຄວາມສຳຄັນຫຼາຍທີ່ສຸດຕໍ່ການໃຫ້ບໍລິການ ເຊິ່ງຈະເປັນການບົ່ງບອກທີ່ສຳຄັນຂອງການໃຫ້ບໍລິການ ແລະ ເປັນຜົນກະທົບໂດຍກົງຕໍ່ລາຍຮັບ ແລະ ກຳໄລໃນການດຳເນີນທຸລະກິດ ໂດຍສ່ວນໃຫ່ຍແລ້ວຈະມີການປະເມີນປະຈຳເດືອນ ຫຼື ປະຈຳປີ. ການຄາດຄະເນການຢຸດໃຊ້ບໍລິການໄດ້ຖືກໃຊ້ໃນທຸກອຸດສາຫະກຳ ໂດຍສະເພາະແມ່ນທາງດ້ານ ໂທລະຄົມມະນາຄົມ ເນື່ອງຈາກວ່າ ມີການໃຊ້ເຕັກໂນໂລຊີການຮຽນຮູ້ ທີ່ມີຄວາມສະຫຼາດ ແລະ ມີປະສິດທິພາບ ເຮັດໃຫ້ສາມາດຄາດຄະເນການເພີ່ມຂຶ້ນ ແລະ ຫຼຸດລົງຂອງລູກຄ້າໄດ້ (Lalwani et al., 2021).

ການຢຸດໃຊ້ບໍລິການ (Customer Churn) ໃນການສື່ສານຈະເກີດຂຶ້ນດ້ວຍຫຼາຍສາເຫດເຊັ່ນ: ຄຸນນະພາບການບໍລິການບໍ່ເປັນທີ່ໜ້າພໍໃຈ, ຄ່າບໍລິການທີ່ບໍ່ເໝາະສົມ ລວມເຖິງບໍລິການລູກຄ້າບໍ່ໄດ້ດີເທົ່າ

ທີ່ຄວນ ຖ້າຢູ່ໃນສະພາບດັ່ງກ່າວນັ້ນ ໜ້າທີ່ສຳຄັນຂອງຜູ້ໃຫ້ບໍລິການຄື: ຕ້ອງເກັບກຳ ແລະ ວິເຄາະຂໍ້ມູນ ເພື່ອເຂົ້າໃຈເຖິງພຶດຕິກຳຂອງລູກຄ້າ ແລ້ວສ້າງກົນໄກການປ້ອງກັນ ເພື່ອບໍ່ໃຫ້ການຢຸດໃຊ້ບໍລິການເກີດຂຶ້ນໄດ້ຢ່າງຕໍ່ເນື່ອງ. ໃນປັດຈຸບັນ ເຕັກໂນໂລຊີການຮຽນຮູ້ ໄດ້ກາຍມາເປັນເຕັກໂນໂລຊີທີ່ຈຳເປັນສຳລັບການຄາດຄະເນການຢຸດໃຊ້ບໍລິການດ້ວຍຄວາມສາມາດໃນການວິເຄາະ ແລະ ຮຽນຮູ້ຈາກຂໍ້ມູນໃນຈຳນວນຫຼາຍໄດ້ຢ່າງມີປະສິດທິພາບ ເພື່ອຊ່ວຍວິເຄາະເຖິງພຶດຕິກຳ ແລະ ຫາຮູບແບບຂອງຂໍ້ມູນທີ່ຈະເປັນຕົວຊີ້ບອກວ່າລູກຄ້າມີແນວໂນ້ມທີ່ຈະອອກຈາກການໃຫ້ບໍລິການສູງ. ເຕັກໂນໂລຊີການຮຽນຮູ້ແມ່ນ ລະບົບການຮຽນຮູ້ຄອມພິວເຕີ ທີ່ກ່ຽວຂ້ອງກັບການພັດທະນາ ແລະ ການສຶກສາສູດຄິດໄລ່ໃນທາງສະຖິຕິ ທີ່ສາມາດຮຽນຮູ້ຈາກຂໍ້ມູນ ໂດຍເປັນຂໍ້ມູນທີ່ສັບຊ້ອນ ແລະ ການເຮັດວຽກທີ່ບໍ່ມີຄຳສັ່ງທີ່ຊັດເຈນ (Samuel, 1986).

ການສຶກສານີ້ໄດ້ດຳເນີນເພື່ອປະເມີນປະສິດທິພາບຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ Logistic Regression (LR) ແມ່ນເຕັກໂນໂລຊີການຮຽນຮູ້ການຫົດຖອຍ ທີ່ຄົ້ນຫາຮູບແບບທີ່ງ່າຍດາຍ ແຕ່ມີປະສິດທິພາບ ທີ່ເໝາະສົມທີ່ສຸດ ໃນການອະທິບາຍຄວາມສຳພັນລະຫວ່າງຕົວປ່ຽນຕາມ ຕາມການຕອບສະໜອງຂອງຕົວປ່ຽນອິດສະຫຼະ (Hosmer & Lemeshow, 2013) ແລະ Random Forest (RF) ແມ່ນເຕັກໂນໂລຊີການຮຽນຮູ້ປ່າໄມ້ແບບສຸ່ມ ເປັນເຕັກໂນໂລຊີການຮຽນຮູ້ເຄື່ອງຈັກຍອດນິຍົມ ທີ່ພັດທະນາໂດຍ Leo Breiman ແລະ Adele Cutler, ໂດຍການເອົາ ເຕັກໂນໂລຊີການຕັດສິນໃຈແບບຕົ້ນໄມ້ ຈຳນວນຫຼາຍມາລວມເຂົ້າກັນເພື່ອໃຫ້ໄດ້ຜົນຮັບອັນດຽວ. ເຕັກໂນໂລຊີການຮຽນຮູ້ນີ້ ມີຄວາມເປັນມິດກັບຜູ້ໃຊ້ ແລະ ຄວາມສະຫຼາດວ່ອງໄວ ເຮັດໃຫ້ເໝາະສົມສຳລັບທັງວຽກງານການຈັດປະເພດການຫົດຖອຍຢ່າງຫຼວງຫຼາຍ (Sruthi, 2024).

ຈາກບັນຫາ ແລະ ຄວາມສຳຄັນທີ່ກ່ວ່າມາຂ້າງເທິງນີ້ ຜູ້ສຶກສາຈຶ່ງມີຄວາມສົນໃຈເພື່ອສຶກສາປະສິດທິພາບຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ ໃນການຄາດຄະເນການໃຊ້ບໍລິການຂອງລູກຄ້າ ບໍລິສັດ TPLUS DIGITAL, ສຶກສາຄວາມໄວໃນການຮຽນຮູ້ ແລະ ຫົດສອບ ຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ ແລະ ເພື່ອປຽບທຽບປະສິດທິພາບ ຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ໃນການຄາດຄະເນການຢຸດໃຊ້ງານຂອງລູກຄ້າ.

2. ອຸປະກອນ ແລະ ວິທີການ

ຮູບແບບວິທີການສຶກສາຄົ້ນຄວ້າແມ່ນໄດ້ອອກແບບແຜນວາດຂັ້ນຕອນວິທີການ ໃນການປະເມີນປະສິດທິພາບຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ ດັ່ງຮູບທີ 1

2.1 ຂໍ້ມູນທີ່ນຳມາວິເຄາະ

ການຄົ້ນຄວ້ານີ້ແມ່ນໃຊ້ຊຸດຂໍ້ມູນ ຈາກຖານຂໍ້ມູນຂອງບໍລິສັດ TPLUS DIGITAL ໃນຮູບແບບນາມສະກຸນ .CSV ມີຂໍ້ມູນທັງຫມົດ 1,003,101 ແຖວ. ໃນການທົດລອງໄດ້ແບ່ງຂໍ້ມູນການຮຽນຮູ້ ແລະ ທົດສອບ ອອກເປັນ 3 ສ່ວນເພື່ອ ການຮຽນຮູ້ ແລະ ການທົດສອບຄື: 70%-30%, 80%-20% ແລະ 90%-10% ຕາມລຳດັບ.

2.2 ປະຊາກອນ ແລະ ກຸ່ມຕົວຢ່າງ

ປະຊາກອນໃນການສຶກສານີ້ແມ່ນລູກຄ້າທີ່ໃຊ້ບໍລິການຂອງບໍລິສັດ TPLUS DIGTTAL ໃນລະບົບຂໍ້ມູນຈິງ ນັບຕັ້ງແຕ່ເດືອນກຸມພາ ຫາ ເດືອນເມສາ ປີ 2024 ທີ່ມີທັງຜູ້ໃຊ້ທີ່ຢຸດໃຊ້ ແລະ ຍັງໃຊ້ງານຢູ່. ຂໍ້ມູນປະກອບໄປດ້ວຍຂໍ້ມູນເລກໝາຍຜູ້ໃຊ້ບໍ່ລິການ, ການເຕີມເງິນ, ການຊື້ແພັກເກັດດາຕ້າ, ການໂທ, ການສົ່ງຂໍ້ຄວາມ, ການໃຊ້ບໍລິການເສີມ, ຍອດເງິນທີ່ໃຊ້ທັງໝົດ, ໄລຍະເວລາການໃຊ້ບໍລິການ, ເພດ, ການຮ້ອງທຸກຂອງລູກຄ້າ ແລະ ປະເພດຂອງລູກຄ້າ ດັ່ງໃນໃນຕາຕະລາງທີ 1 ສະແດງໃຫ້ເຫັນເຖິງລາຍລະອຽດຂໍ້ມູນທີ່ມີຊື່ຕົວປ່ຽນ ແລະ ລາຍລະອຽດຂອງຕົວປ່ຽນ.

2.3 ການກະກຽມຂໍ້ມູນ (Data Preprocessing)

2.3.1 ການທຳຄວາມສະອາດຂໍ້ມູນ (Data Cleaning)

ຄຸນນະພາບຂອງຂໍ້ມູນເປັນປັດໃຈຫຼັກຕໍ່ປະສິດທິພາບ ຂອງເຕັກໂນໂລຊີການຮຽນຮູ້. ໃນການສຶກສານີ້ ມີການດຳເນີນຂະບວນການທຳຄວາມສະອາດຂໍ້ມູນ, ການຈັດການຂໍ້ມູນທີ່ສູນຫາຍ, ຂໍ້ມູນທີ່ຊ້ຳກັນ ແລະ ແກ້ໄຂຂໍ້ມູນທີ່ບໍ່ຖືກຕ້ອງ ເພື່ອຮັບປະກັນໃນຄວາມໜ້າເຊື່ອຖື ແລະ ຄວາມຖືກຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ (Fujo et al., 2022).

2.3.2 ການແປງຮູບແບບຂໍ້ມູນ (Data Transformation)

ການປ່ຽນແປງຂໍ້ມູນຈາກຮູບແບບປະເພດໝວດໝູ່ ໃຫ້ກາຍເປັນຕົວເລກ ທີ່ເອີ້ນວ່າ “ການເຂົ້າລະຫັດປະເພດໝວດໝູ່” ແມ່ນຂັ້ນຕອນທີ່ມີຄວາມສຳຄັນຫຼາຍໃນການກະກຽມຂໍ້ມູນ ເຊິ່ງຂະບວນການນີ້ມີຄວາມຈຳເປັນເນື່ອງຈາກວ່າເຮັດໃຫ້ເຕັກໂນໂລຊີການຮຽນຮູ້ມີປະສິດທິພາບຫຼາຍຂຶ້ນ ເມື່ອໃຊ້ຂໍ້ມູນທີ່ເປັນຕົວເລກ (Hall, 1999).

2.3.3 ການປັບປຸງຂໍ້ມູນ (Data Adjustment)

ກ່ອນທີ່ຈະເຮັດການຮຽນຮູ້ ແລະ ທົດສອບຂໍ້ມູນ ເຮົາໄດ້ມີການປັບຄ່າຂອງຊຸດຂໍ້ມູນໃຫ້ມີຄ່າສະເລ່ຍເປັນມາດຕະຖານ ຂັ້ນຕອນນີ້ມີຈຸດປະສົງເພື່ອໃຫ້ຂໍ້ມູນຢູ່ໃນມາດຕະຖານດຽວກັນ ເຊິ່ງການປັບຄ່າແບບນີ້ຈະຊ່ວຍໃຫ້ເຕັກໂນໂລຊີເຮັດວຽກມີປະສິດທິພາບ ແລະ ຫຼຸດການລຳອຽງຈາກການທີ່ຂໍ້ມູນມີຄ່າຫຼາຍ ຫຼື ມີສ່ວນຫຼົ່ນກັນຫຼາຍເກີນໄປ (Imani et al., 2023).

2.4 ການເລືອກຄຸນສົມບັດຂໍ້ມູນ (Feature Selection)

ວິທີການເລືອກຄຸນລັກສະນະ ທີ່ອີງໃສ່ສະຖິຕິ ແມ່ນວິທີການເຊື່ອມໂຍງລະຫວ່າງຂໍ້ມູນ ແຕ່ລະຄ່າຄຸນລັກສະນະທີ່ໃຊ້ເປັນຕົວປ່ຽນນຳເຂົ້າ ກັບຜົນຮັບທີ່ຕ້ອງການໃນການຄາດຄະເນ ໂດຍໃຊ້ວິທີການວິເຄາະທາງສະຖິຕິ. ຕົວປ່ຽນນຳເຂົ້າທີ່ມີຄວາມສຳພັນຢ່າງສູງ ຫຼື ມີຜົນຕໍ່ເປົ້າໝາຍແບບຊັດເຈນ ແມ່ນຕົວປ່ຽນທີ່ຖືກເລືອກໃຊ້ກັບເຕັກໂນໂລຊີ ເພື່ອປັບປຸງຄວາມຖືກຕ້ອງ ແລະ ປະສິດທິພາບຂອງການຄາດຄະເນ (Kavitha, 2020).

2.5 ການແຍກຂໍ້ມູນໃນການຮຽນຮູ້ ແລະ ທົດສອບ

ການແບ່ງຊຸດຂໍ້ມູນເພື່ອຮຽນຮູ້ ແລະ ທົດສອບ ໂດຍວິທີແຍກແບບຊຸ່ມ ເປັນວິທີທີ່ໄດ້ຮັບການຍອມຮັບຢ່າງກວ້າງຂວາງໃນຂົງເຂດການຄົ້ນຄວ້າ ແລະ ວິໄຈ ດ້ານເຕັກໂນໂລຊີການຮຽນຮູ້ ມາແຕ່ດົນນານ (Kosaraju et al., 2023). ວັດຖຸປະສົງຫຼັກຂອງວິທີນີ້ຄືການປະເມີນຄວາມຖືກຕ້ອງໃນການຄາດຄະເນຂອງເຕັກໂນໂລຊີ ວ່າຈະປ່ຽນແປງແນວໃດເມື່ອຖືກຮຽນຮູ້ກັບຂໍ້ມູນໃໝ່ ແລະ ຖືກທົດສອບກັບຂໍ້ມູນທີ່ບໍ່ເຄີຍເຫັນມາກ່ອນ (Li et al., 2021).

ໃນງານຄົ້ນຄວ້ານີ້ໄດ້ແບ່ງຊຸດຂໍ້ມູນອອກເປັນ 70%-30%, 80%-20% ແລະ 90%-10% ດ້ວຍວິທີການແຍກແບບຊຸ່ມ. ວິທີນີ້ຊ່ວຍໃຫ້ໝັ້ນໃຈວ່າ ການແບ່ງຂໍ້ມູນເປັນໄປຢ່າງຍຸດຕິທຳໃນຂະບວນການທົດສອບ ເພື່ອໃຫ້ການປະເມີນປະສິດທິພາບຂອງເຕັກໂນໂລຊີເປັນໄປໄດ້ຢ່າງຖືກຕ້ອງໃນເງື່ອນໄຂທີ່ສະໝ່ຳສະເໝີ.

2.6 ຊຸດຂໍ້ມູນທີ່ເປັນຕົວແທນ ແລະ ອະຄະຕິ (Dataset Representativeness and Biases)

ໃນຂະບວນການອອກແບບ ແລະ ພັດທະນາເຕັກໂນໂລຊີການຮຽນຮູ້ ການມີຊຸດຂໍ້ມູນທີ່ມີຄວາມເປັນຕົວແທນຂອງຄວາມເປັນຈິງ ແມ່ນຫົວໃຈສຳຄັນໃນການຮັບປະກັນວ່າເຕັກໂນໂລຊີທີ່ຖືກຮຽນຮູ້ ຈະມີຄວາມສາມາດໃນການຄາດຄະເນໄດ້ຢ່າງຖືກຕ້ອງໃນສະພາບການທີ່ເປັນຈິງ. ຖ້າຂໍ້ມູນຖືກເກັບກຳມາຈາກແຫຼ່ງທີ່ບໍ່ຄອບຄຸມ ຫຼື ມີການລຳອຽງໃນການຄັດເລືອກຕົວຢ່າງ ເຕັກໂນໂລຊີທີ່ໄດ້ຮັບການຮຽນນັ້ນອາດຈະຮຽນຮູ້ພຶດຕິກຳທີ່ບໍ່ຖືກຕ້ອງ ແລະ ນຳໄປສູ່ການຄາດຄະເນທີ່ບໍ່ຖືກຕ້ອງໃນສະພາບແວດລ້ອມຈິງ.

ເພື່ອຫຼຸດຜົນກະທົບຈາກຂໍ້ມູນທີ່ເປັນອະຄະຕິ ຕ້ອງວິເຄາະຂໍ້ມູນໃນສ່ວນຂອງກຸ່ມຕ່າງໆ ຄວາມຫຼາກຫຼາຍຂອງຕົວຢ່າງ ການກະຈາຍຂອງພຶດຕິກຳທີ່ສະທ້ອນຄວາມເປັນຈິງ; ການໃຊ້ຂໍ້ມູນທີ່ເປັນອະຄະຕິສາມາດເກີດຂຶ້ນໄດ້ຈາກ ຄວາມລຳອຽງໃນຂະບວນການເກັບຂໍ້ມູນ, ຄວາມບໍ່ສະເໝີພາບທາງປະຫວັດສາດ ຫຼື ການຂາດສິດທິດເທົ່າທຽມ, ການຄັດເລືອກຕົວຢ່າງທີ່ ຂາດຂໍ້ມູນຈາກກຸ່ມທີ່ສຳຄັນ (Shahbazi et al., 2023).

ດັ່ງນັ້ນ ການປະກັນຄຸນນະພາບຂອງຊຸດຂໍ້ມູນຈຶ່ງເປັນຫົວໃຈຂອງຄວາມສຳຄັນໃນການປະຍຸກໃຊ້ເຕັກໂນໂລຊີການຮຽນຮູ້ໃນສະພາບຄວາມເປັນຈິງ.

2.7 ການກວດສອບ

ໃນຂະບວນການພັດທະນາເຕັກໂນໂລຊີການຮຽນຮູ້ ການກວດສອບແມ່ນ ຂັ້ນຕອນທີ່ມີບົດບາດສຳຄັນໃນການປະເມີນປະສິດທິພາບ ແລະ ການທົດສອບຄວາມແນ່ນອນຂອງເຕັກໂນໂລຊີທີ່ຖືກຮຽນຮູ້ຈາກຊຸດຂໍ້ມູນ ແລ້ວໄປທົດສອບກັບຂໍ້ມູນທີ່ບໍ່ໄດ້ນຳໃຊ້ໃນການຮຽນຮູ້ໂດຍກົງ. ຈຸດປະສົງຫຼັກຂອງການກວດສອບ ບໍ່ໃຊ້ພຽງແຕ່ເພື່ອຄັດເລືອກເຕັກໂນໂລຊີທີ່ດີທີ່ສຸດ ຫຼື ປັບຄ່າໃຫ້ເໝາະສົມເທົ່ານັ້ນ ແຕ່ຍັງຊ່ວຍປະເມີນຄວາມສາມາດໃນການຄາດຄະເນຂອງເຕັກໂນໂລຊີຕໍ່ຂໍ້ມູນໃໝ່ໄດ້ຢ່າງຖືກຕ້ອງ. ຂໍ້ມູນທີ່ນຳໃຊ້ໃນຂະບວນການຮຽນຮູ້ນັ້ນມີບົດບາດສຳຄັນເທົ່າກັບໂຄງສ້າງ ແລະ ກະບວນການຮຽນຮູ້, ການແຍກຊຸດຂໍ້ມູນເພື່ອທົດສອບຢ່າງຖືກຕ້ອງ ຈຶ່ງເປັນສິ່ງຈຳເປັນໃນການປະກັນຄວາມເຊື່ອຖື ແລະ ປະສິດທິພາບຂອງເຕັກໂນໂລຊີ (Paleyes et al., 2022).

2.7.1 ການວັດຜົນການປະເມີນເຕັກໂນໂລຊີດ້ວຍ Confusion Matrix

ເປັນເຄື່ອງມືສຳລັບການວິເຄາະການຄາດຄະເນ ໃນເຕັກໂນໂລຊີການຮຽນຮູ້ ເພື່ອກວດສອບໃນການຄາດຄະເນໃນຮູບແບບຕາຕະລາງຄວາມສຳພັນຂອງຂໍ້ມູນ ທີ່ມີຈຳນວນການຄາດຄະເນທີ່ຖືກຕ້ອງ ແລະ ບໍ່ຖືກຕ້ອງໃນຮູບແບບການຈັດປະເພດ (Karimi, 2021) ດັ່ງທີ່ສະແດງໃນຕາຕະລາງທີ 2.

- True Positive (TP) ການຄາດຄະເນບໍ່ຢຸດໃຊ້ ຄ່າຄວາມຈິງບໍ່ຢຸດໃຊ້ (ຖືກ).
- True Negative (TN) ການຄາດຄະເນຈະຢຸດໃຊ້ ຄ່າຄວາມຈິງຢຸດໃຊ້ (ຖືກ).
- False Positive (FP) ການຄາດຄາດຄະເນບໍ່ຢຸດໃຊ້ ຄ່າຄວາມຈິງຢຸດໃຊ້ (ຜິດ).
- False Negative (FN) ການຄາດຄະເນຢຸດໃຊ້ ແຕ່ຄ່າຄວາມຈິງບໍ່ຢຸດໃຊ້ (ຜິດ).

2.7.2 ຕົວຊີ້ວັດໃນການປະເມີນ

ໃນກໍລະນີທີ່ຊຸດຂໍ້ມູນມີບັນຫາຄວາມລຳອຽງຂອງການແຍກປະເພດ ເຊັ່ນ: ຈຳນວນລູກຄ້າທີ່ຈະຢຸດໃຊ້ບໍລິການມີໜ້ອຍກວ່າຜູ້ທີ່ບໍ່ໄດ້ຢຸດໃຊ້. ການນຳຄ່າຄວາມຖືກຕ້ອງໄປໃຊ້ປະເມີນຜົນໂດຍລວມອາດຈະບໍ່ສິ່ງຜົນສະທ້ອນຄວາມຖືກຕ້ອງໄດ້ຢ່າງແທ້ຈິງ. ຍົກຕົວຢ່າງ: ຖ້າຂໍ້ມູນ 95% ແມ່ນການໃຊ້ງານ ແລະ ເຕັກໂນໂລຊີຄາດຄະເນວ່າລູກຄ້າທຸກຄົນແມ່ນບໍ່ຢຸດໃຊ້ງານ ກໍ່ຈະໄດ້ຄ່າຄວາມຖືກຕ້ອງສູງ ແຕ່ຂາດຄວາມສາມາດໃນການຈັບກຸ່ມລູກຄ້າຜູ້ທີ່ຢຸດໃຊ້ງານ ທີ່ມີຄວາມສຳຄັນທາງທຸລະກິດ.

ດັ່ງນັ້ນ ຄ່າວັດທີ່ເໝາະສົມຫຼາຍກວ່າໃນສະຖານະການນີ້ມີຄື:

- Accuracy: ຄ່າຄວາມຖືກຕ້ອງ
- Recall (Sensitivity): ຄ່າຄວາມສາມາດໃນການຈັບກຸ່ມລູກຄ້າທີ່ຢຸດການນຳໃຊ້ໄດ້ຖືກຕ້ອງ (True Positive Rate)
- Precision: ຈຳນວນການຄາດຄະເນທີ່ເປັນຈິງທີ່ຖືກ
- F1-score: ຄ່າປະສິດທິດພາບຄວາມຖືກຕ້ອງ
- AUC-ROC (Area Under the Curve - Receiver Operating Characteristic): ແມ່ນຄ່າຄວາມສາມາດໃນການຄາດຄະເນລູກຄ້າທີ່ຈະຢຸດ ແລະ ບໍ່ຢຸດໃຊ້ງານ

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

2.7.3 ການປະເມີນ

ໃນຂະບວນການປະເມີນຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ ໄດ້ໃຊ້ປະສິດທິພາບໃນການແຍກປະເພດ (ROC) Curve ຢູ່ໃນພື້ນທີ່ (AUC). ROC Curve ແມ່ນຮູບແບບການສະແດງຜາບສຳລັບເຕັກໂນໂລຊີການຮຽນຮູ້ ສາມາດຈຳແນກໄດ້ 2 ຄລາສ (Binary Classifier) ໂດຍມີການສະແດງຄວາມສຳພັນລະຫວ່າງ ອັດຕາການຄາດຄະເນຄ່າຄວາມຈິງທີ່ຖືກ (TPR) ແລະ ອັດຕາການຄາດຄະເນຄ່າຄວາມຈິງທີ່ຜິດ (FPR) ທີ່ເກີດຂຶ້ນໃນພື້ນທີ່ຈຸດຕັດ ທີ່ແຕກຕ່າງກັນ (Saypadith & Onoye, 2021).

-True Positive Rate

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

- False Positive Rate

$$FPR = \frac{TP}{FP + TN} \quad (6)$$

2.8 ການຄາດຄະເນ ແລະ ການແຍກປະເພດ

ການຄາດຄະເນການຢຸດໃຊ້ບໍລິການ (Churn Prediction) ແມ່ນໜຶ່ງໃນແນວທາງທີ່ສຳຄັນສຳລັບຂົງເຂດທຸລະກິດໂທລະຄົມມະນາຄົມທີ່ເປົ້າໝາຍຫຼັກ ແມ່ນການປະເມີນຫາລູກຄ້າຜູ້ທີ່ຈະຢຸດໃຊ້ບໍລິການ ລວມເຖິງການພັດທະນາຍຸດທະສາດໃນການຮັກສາລູກຄ້າໃຫ້ໄດ້ຜົນດີທີ່ສຸດ. ຫຼາຍບົດຄົ້ນຄວ້າໄດ້ສະແດງໃຫ້ເຫັນວ່າ ມີຫຼາຍວິທີການທີ່ສາມາດນຳມາໃຊ້ໃນການຄາດຄະເນເຊັ່ນ: ວິທີການສະຖິຕິດັ່ງເດີມ ແລະ ເຕັກນິກການຮຽນຮູ້ທັນສະໄໝ. ໃນການຄາດຄະເນການຢຸດຂອງລູກຄ້າແມ່ນມີຄຸນຄ່າຢ່າງຫຼວງຫຼາຍໃນການວາງແຜນຍຸດທະສາດຂອງບໍລິສັດ ເພາະຊ່ວຍໃນການຈັບກຸ່ມລູກຄ້າທີ່ກໍ່ໃຫ້ເກີດຄວາມສູນເສຍລາຍໄດ້ ແລະ ສາມາດວາງມາດຕະການເພື່ອຮັກສາໄວ້ໄດ້ (Hashmi et al., 2013).

2.8.1 Random Forest

ແມ່ນວິທີການຮຽນຮູ້ຂອງເຕັກໂນໂລຊີຊະນິດໜຶ່ງ ທີ່ສ້າງຂຶ້ນຈາກການລວມກັນຂອງຫຼາຍຮູບແບບການຕັດສິນໃຈແບບຕົ້ນໄມ້ແບບສຸ່ມ ໂດຍໃນຂະບວນການຄາດຄະເນ ແມ່ນໃຊ້ຫຼັກການໃຫ້ຄຳຕອບທີ່ຄືກັນຫຼາຍສຸດ ສຳລັບການຈັດປະເພດ ຫຼື ສຳລັບຄ່າຕໍ່ເນື່ອງ (Biau & Scornet, 2016). ຈຸດເດັ່ນຂອງ RF ຄື: ມີຄວາມສາມາດໃນການຮັບມືກັບຂໍ້ມູນຂະໜາດໃຫ່ຍ ແລະ ຊຸດຂໍ້ມູນທີ່ມີຄວາມສັບຊ້ອນ, ສາມາດຈັດການກັບຄວາມສຳພັນ ລະຫວ່າງຄຸນລັກສະນະຂໍ້ມູນໄດ້ດີ ໂດຍບໍ່ຈຳເປັນຕ້ອງປັບຄ່າກໍ່ໄດ້, ຫຼຸດຜ່ອນຄວາມເສຍຫາຍຈາກການຮຽນຮູ້ຫຼາຍເກີນໄປໄດ້ດີ ເພາະໃຊ້ການສຸ່ມແຍກຊຸດຂໍ້ມູນຍ່ອຍໄວ້ສຳລັບການທົດສອບ. ດັ່ງໃນຮູບທີ 2 ແມ່ນຮູບແບບການເຮັດວຽກຂອງ Random Forest.

2.8.2 Logistic Regression

ແມ່ນວິທີການວິເຄາະທາງສະຖິຕິທີ່ນິຍົມໃຊ້ກັນຢ່າງກວ້າງຂວາງໃນການແຍກປະເພດ ໂດຍພິເສດໃນດ້ານການຄາດຄະເນຜູ້ໃຊ້ທີ່ມີແນວໂນ້ມຈະຢຸດໃຊ້ບໍລິການ ວິທີການນີ້ໃຊ້ໃນການປະເມີນຄວາມອາດຈະເປັນໄປໄດ້ ທີ່ຫຼຸດລົງມາໃນຊ່ວງ 0 ຫາ 1 ຜ່ານຟັງຊັນ Sigmoid ເພື່ອຊີ້ວ່າລູກຄ້າຄົນໜຶ່ງມີແນວໂນ້ມຈະຢຸດໃຊ້ບໍລິການຫຼາຍໜ້ອຍເທົ່າໃດ. ຈຸດແຂງຂອງ LR ຄືມີຄວາມງ່າຍດາຍໃນການແຍກປະເພດ, ມີປະສິດທິພາບທີ່ດີໃນຂໍ້ມູນທີ່ບໍ່ຊັບຊ້ອນ ເຊິ່ງເປັນພື້ນຖານສຳຄັນໃນການປຽບທຽບກັບເຕັກໂນໂລຊີອື່ນ. LR ແມ່ນເຄື່ອງມືທີ່ມີຄວາມເປັນລະບຽບ ແລະ ເໝາະສົມຕໍ່ວຽກງານທີ່ຕ້ອງການຄວາມຖືກຕ້ອງ ແລະ ຄວາມເຂົ້າໃຈງ່າຍໃນຜົນການວິເຄາະຂອງເຕັກໂນໂລຊີ (Tan et al., 2021). ດັ່ງໃນຮູບທີ 3 ແມ່ນ ຮູບແບບການເຮັດວຽກຂອງ Logistic Regression ໂດຍໃຊ້ Sigmoid Function.

3. ຜົນໄດ້ຮັບ

ຈາກການທົດສອບ ແລະ ປຽບທຽບລະຫວ່າງ LR ແລະ RF ໃນການຄາດຄະເນການຢຸດໃຊ້ບໍລິການ ສາມາດສະຫຼຸບໄດ້ ດັ່ງນີ້:

- ຈາກຜົນການທົດລອງດ້ານຄວາມໄວໃນການຮຽນຮູ້ ໃນຕາຕະລາງທີ 3 ແລະ ການທົດສອບ ໃນຕາຕະລາງທີ 4 ສະແດງໃຫ້ເຫັນວ່າ LR ໂດຍມີຄວາມໄວກວ່າ RF.
- ຈາກຜົນການທົດລອງ ດ້ານຄວາມສາມາດໃນການຄາດຄະເນລູກຄ້າທີ່ຈະຢຸດ ແລະ ບໍ່ຢຸດໃຊ້ງານ ໃນຕາຕະລາງທີ 5 ສະແດງໃຫ້ເຫັນວ່າ RF ມີຄ່າທີ່ດີກວ່າ LR ເຊັ່ນດຽວກັນ.
- ຈາກຜົນການທົດລອງ ການປະເມີນປະສິດທິພາບ ຂອງເຕັກໂນໂລຊີ ໃນຕາຕະລາງທີ 6 ສະແດງໃຫ້ເຫັນວ່າ LR ແລະ RF ສາມາດຮັບມືກັບຄວາມແຕກຕ່າງຂອງຂໍ້ມູນການຮຽນຮູ້ໄດ້ເປັນຢ່າງດີ. ຕົວຢ່າງ ການຮຽນຮູ້ ໃນສັດສ່ວນຂໍ້ມູນທີ່ຫຼາຍຂຶ້ນ ຈະເຫັນວ່າໃຊ້ຊຸດຂໍ້ມູນຮຽນຮູ້ຫຼາຍຂຶ້ນ ແລະ ມີຄ່າຄວາມສາມາດໃນການຈັບກຸ່ມລູກຄ້າທີ່ຢຸດການນຳໃຊ້ໄດ້ຖືກຕ້ອງ ແລະ ຄ່າປະສິດທິດພາບຄວາມຖືກຕ້ອງສູງຂຶ້ນ ໂດຍສະເພາະ RF ທີ່ສະແດງໃຫ້ເຫັນຄວາມສາມາດໃນການຮຽນຮູ້ຈາກຂໍ້ມູນທີ່ມີປະລິມານຫຼາຍ.

ຜົນຂອງ Confusion Matrix ຈາກຕາຕະລາງທີ 7 ສະແດງວ່າ RF ສາມາດຄາດຄະເນລູກຄ້າທີ່ຈະຢຸດໃຊ້ບໍລິການໄດ້ຖືກຕ້ອງຫຼາຍກວ່າ ເພາະ LR ມີຄ່າການຄາດຄະເນຜິດທີ່ສູງກວ່າ.

ດັ່ງນັ້ນ ຈາກຜົນການທົດລອງຈຶ່ງສາມາດເວົ້າໄດ້ວ່າ RF ເໝາະສົມສຳລັບການນຳໃຊ້ເປັນເຕັກໂນໂລຊີຫຼັກ ໃນການຄາດຄະເນການຢຸດໃຊ້ບໍລິການ ແລະ LR ສາມາດນຳໃຊ້ໄດ້ໃນກໍລະນີທີ່ຕ້ອງການຄວາມໄວ ແລະ ຊໍ້ມູນບໍ່ສັບຊ້ອນ.

4. ວິພາກຜົນ

ໃນການທົດສອບເຕັກໂນໂລຊີການຮຽນຮູ້ LR ທີ່ໃຊ້ສຳລັບການທົດສອບການຄາດຄະເນລູກຄ້າທີ່ອາດຈະຢຸດໃຊ້ບໍລິການ. ຜົນການທົດສອບທຸກການແບ່ງສ່ວນຂໍ້ມູນ ໃນການຮຽນຮູ້ ແລະ ທົດສອບ 70%-30%, 80%-20% ແລະ 90%-10% ສະແດງໃຫ້ເຫັນວ່າ:

- ຜົນການທົດລອງຄວາມໄວຂອງ LR ໃນສັດສ່ວນການແບ່ງຂໍ້ມູນ 70%, 80%, 90% ໃນການຮຽນຮູ້ແມ່ນມີຄວາມໄວກວ່າ RF ຫຼາຍກວ່າ 3000%
- ຜົນການທົດລອງໃນການວັດຄ່າຄວາມຖືກຕ້ອງໃນການຄາດຄະເນ ແມ່ນຢູ່ໃນລະດັບດີ ແຕ່ວ່າມີຄ່ານ້ອຍກວ່າ RF 1.25%, ຄ່າຄວາມສາມາດໃນການຈັດກຸ່ມລູກຄ້າທີ່ຖືກ ນ້ອຍກວ່າ 1.14% ແລະ ຄ່າປະສິດທິດພາບຄວາມຖືກຕ້ອງ ນ້ອຍກວ່າ 2.30% ເມື່ອປຽບທຽບກັບ RF.

ຈາກຜົນການທົດລອງນີ້ສອດຄ່ອງກັບ (Hosmer & Lemeshow, 2013) ທີ່ກ່າວໄວ້ວ່າ LR ເໝາະສົມສຳລັບປັນຫາຄາດຄະເນສອງຄ່າ ແຕ່ອາດຂາດຄວາມຍືດຫຍຸ່ນໃນຂໍ້ມູນສັບຊ້ອນ.

ໃນການທົດສອບເຕັກໂນໂລຊີການຮຽນຮູ້ RF ແມ່ນເຕັກໂນໂລຊີການຮຽນຮູ້ ທີ່ໃຊ້ຢ່າງແຜ່ຫຼາຍໃນການ Churn Prediction ສະແດງໃຫ້ເຫັນວ່າ:

- ຈາກຜົນການທົດລອງ ສະແດງໃຫ້ເຫັນວ່າ ຄ່າຄວາມສາມາດໃນການຈັດກຸ່ມລູກຄ້າທີ່ຖືກ, ຄ່າປະສິດທິດພາບຄວາມຖືກຕ້ອງ ແລະ ຄວາມຖືກຕ້ອງໃນການຄາດຄະເນຫາລູກຄ້າຜູ້ທີ່ຈະຢຸດໃຊ້ງານ ຂອງ RF ແມ່ນມີຄ່າ ສູງກວ່າ LR
- ຈາກຜົນການທົດລອງຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ ກັບຂໍ້ມູນທີ່ສັບຊ້ອນ ສະແດງໃຫ້ເຫັນວ່າມີການຈັດການຄວາມສຳພັນຂອງຄຸນລັກຊະນະຂອງຂໍ້ມູນໄດ້ດີໄດ້ດີ ແລະ ຄ່າໃນການຄາດຄະເນຄວາມເປັນຈິງທີ່ຖືກກໍເປັນທີ່ໜ້າພໍໃຈ ເມື່ອປຽບທຽບກັບ LR ແມ່ນຫລາຍກວ່າ 1.15%

ຈາກຜົນການທົດລອງນີ້ສອດຄ່ອງກັບ (Lalwani et al., 2021) ແລະ (Fujo et al., 2022) ທີ່ສະແດງວ່າ RF ເປັນເຕັກໂນໂລຊີການຮຽນຮູ້ ທີ່ດີທີ່ສຸດໃນການທົດສອບໂຄງສ້າງທີ່ຊັບຊ້ອນຂອງຊຸດຂໍ້ມູນ.

ຜົນການແບ່ງສັດສ່ວນຊຸດຂໍ້ມູນ 70%-30%, 80%-20%, 90%-10% ໃນການຮຽນຮູ້ ແລະ ທົດສອບ ມີຜົນຕໍ່ຄວາມຖືກຕ້ອງ ແລະ ຄວາມສົມດຸນຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ດັ່ງ ລຸ່ມນີ້:

- ໃນການແບ່ງສັດສ່ວນຊຸດຂໍ້ມູນໃນການຮຽນຮູ້ ແລະ ທົດສອບ 70%-30% ເປັນຄ່າມາດຕະຖານ ທີ່ໃຫ້ຜົນທົດສອບທີ່ໜ້າເຊື່ອຖື

- ໃນການແບ່ງສັດສ່ວນຊຸດຂໍ້ມູນໃນການຮຽນຮູ້ ແລະ ທົດສອບ 90%-10% ຈະສົ່ງຜົນໃຫ້ມີການຮຽນຮູ້ຫຼາຍເກີນໄປສຳລັບ RF ເນື່ອງຈາກຂໍ້ມູນ ທົດສອບໜ້ອຍ

- ໃນການແບ່ງສັດສ່ວນຊຸດຂໍ້ມູນໃນການຮຽນຮູ້ ແລະ ທົດສອບ 80%-20% ແມ່ນມີຄວາມສົມດຸນດີທີ່ສຸດສຳລັບການເຮັດ ບົດວິໄຈໃນຄັ້ງນີ້

ຈາກຜົນການທົດລອງຂອງທ່ານ (Tan et al., 2021) ທີ່ ສຶກສາຜົນຂອງ train/test split ໃນເຕັກໂນໂລຊີການຮຽນຮູ້ໄດ້ ສອດຄ່ອງຕໍ່ຜົນທີ່ໄດ້ຮັບໃນການສຶກສາໃນຄັ້ງນີ້ສະແດງໃຫ້ເຫັນວ່າ: ການນຳໃຊ້ເຕັກໂນໂລຊີການຮຽນຮູ້ RF ໄປປະຍຸກໃຊ້ໃນແພລດຟອມ ການຕິດຕາມລູກຄ້າອາດຈະເຮັດໃຫ້ບໍລິສັດສາມາດແກ້ໄຂການຢຸດໃຊ້ ບໍລິການຂອງລູກຄ້າລ່ວງໜ້າໄດ້. ເປັນພື້ນຖານທີ່ຈະນຳໄປສູ່ການສ້າງ ຮູບແບບ ຍຸດທະສາດໃນການຮັກສາລູກຄ້າເກົ່າ ທີ່ມີປະສິດທິພາບ ແລະ ສ້າງຄວາມພໍໃຈໃນສິນຄ້າ ຫຼື ການໃຫ້ບໍລິການ ໃນໄລຍະຍາວໄດ້.

5. ສະຫຼຸບ

ການສຶກສານີ້ໄດ້ດຳເນີນການປຽບທຽບປະສິດທິພາບຂອງ ສອງເຕັກໂນໂລຊີການຮຽນຮູ້ທີ່ນິຍົມໃຊ້ຢ່າງແຜ່ຫຼາຍ ຄື Logistic Regression (LR) ແລະ Random Forest (RF) ສຳລັບການຄາດ ຄະເນລູກຄ້າທີ່ມີແນວໂນ້ມຈະຢຸດໃຊ້ບໍລິການຂອງບໍລິສັດ TPLUS DIGITAL ຜົນການສຶກສາສະແດງໃຫ້ເຫັນວ່າ: RF ແມ່ນເຕັກໂນໂລ ຊີການຮຽນຮູ້ທີ່ສາມາດຄາດຄະເນໄດ້ຖືກຕ້ອງຫຼາຍກວ່າ LR ໃນຫຼາຍ ມິຕິ ໂດຍສະເພາະ ຄ່າຄວາມສາມາດໃນການຈັດກຸ່ມລູກຄ້າທີ່ຖືກ, ຄ່າ ປະສິດທິຜາບຄວາມຖືກຕ້ອງ ແລະ ຄວາມຖືກຕ້ອງໃນການຄາດຄະເນ ຫາລູກຄ້າຜູ້ທີ່ມີແນວໂນ້ມທີ່ຈະຢຸດໃຊ້ງານໄດ້ແນ່ນອນ. ສຳລັບ LR ແມ່ນເຕັກໂນໂລຊີການຮຽນຮູ້ທີ່ມີຄວາມໄວໃນການຮຽນຮູ້ ແລະ ສາມາດຈັດການກັບຊຸດຂໍ້ມູນທີ່ບໍ່ສົມດຸນໄດ້ດີ ແຕ່ມັກຜິດພາດໃນການ ຄາດຄະເນລູກຄ້າຜູ້ໃຊ້ທີ່ຈະຢຸດໃຊ້ແຕ່ໃນຄວາມເປັນຈິງ ບໍ່ໄດ້ຢຸດໃຊ້ ງານ. ໃນດ້ານ Confusion Matrix ສະແດງໃຫ້ເຫັນວ່າ RF ສາມາດ ຄວບຄຸມຜົນການຈຳແນກໄດ້ດີ ມີຄ່າ ການຄາດຄະເນຢຸດໃຊ້ ແຕ່ຄ່າ ຄວາມຈິງບໍ່ຢຸດໃຊ້ ນ້ອຍກວ່າ LR ໄດ້ຢ່າງຊັດເຈນ ເຊິ່ງແມ່ນສ່ວນສຳ ຄັນໃນການລະວັງການສູນເສຍລູກຄ້າແບບທີ່ບໍ່ຄາດຄິດ. ດັ່ງນັ້ນ ການນຳໃຊ້ RF ໃນການຄາດຄະເນວ່າ ລູກຄ້າຈະຢຸດໃຊ້ບໍລິການ ແມ່ນ ເຕັກໂນໂລຊີການຮຽນຮູ້ທີ່ມີປະສິດທິພາບ ແລະ ເໝາະສົມໃນການນຳ ໄປໃຊ້ວຽກງານທີ່ຕ້ອງການການຄາດຄະເນທີ່ແນ່ນອນ ເພື່ອຈະຮັກສາ ລູກຄ້າໄວ້ໃຫ້ໄດ້ຫຼາຍທີ່ສຸດ. ແຜນການພັດທະນາ ແລະ ຂະຫຍາຍຜົນ ການຄົ້ນຄ້ວາໃນອະນາຄົດ ແມ່ນເພື່ອເພີ່ມປະສິດທິພາບໃຫ້ດີຍິ່ງຂຶ້ນ ສາມາດທົດລອງ ແລະ ປຽບທຽບໃນຮູບແບບໃໝ່ໆ, ນຳເອົາການສ້າງ ຄຸນລັກສະນະຂໍ້ມູນທາງດ້ານວິສະວະກຳຂັ້ນສູງມາຊ່ວຍ, ໃຊ້ຂໍ້ມູນໃນ ເວລາຈິງມາຄາດຄະເນ ແລະ ການຜັນຂະຫຍາຍໃຫ້ເປັນເຕັກໂນໂລຊີ ໃໝ່ທີ່ສ້າງປະໂຫຍດສູງສຸດໃຫ້ແກ່ອົງກອນ, ນັກຮຽນ, ນັກຄົ້ນຄວ້າ ໃນ ຕໍ່ໜ້າ.

6. ຂໍ້ຂັດແຍ່ງ

ຕໍ່ຜົນຄົ້ນຄວ້າ-ວິໄຈ ຂ້າງເທິງທັ້ງໝົດ, ຂ້າພະເຈົ້າໃນນາມຜູ້ ຄົ້ນຄວ້າວິທະຍາສາດ ຂໍປະຕິຍານວ່າ: ຂໍ້ມູນທັ້ງໝົດທີ່ມີ ໃນບົດຄວາມ ວິຊາການດັ່ງກ່າວນີ້ ແມ່ນບໍ່ ມີຂໍ້ ຂັດແຍ່ງທາງຜົນປະໂຫຍດກັບ ພາກສ່ວນໃດ ແລະ ບໍ່ໄດ້ເອື້ອປະໂຫຍດໃຫ້ກັບ ພາກສ່ວນໃດໜຶ່ງ ຫຼື ຜົນປະໂຫຍດຂອງບຸກຄົນໃດໜຶ່ງ. ໃນກໍລະນີມີການລະເມີດໃນຮູບການ ໃດໜຶ່ງ ຂ້າພະເຈົ້າມີຄວາມ ຍິນດີຮັບຜິດຊອບແຕ່ຜຽງຜູ້ດຽວ.

7. ເອກະສານອ້າງອີງ

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Machine learning algorithms for customer churn prediction: A review. *International Journal of Computer Applications*, *184*(7), 1–6.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato.

Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer churn prediction in telecom using machine learning. *International Journal of Computer Applications*, *80*(17), 32–37.

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Imani, M., Mehdi, M., & Arabnia, H. R. (2023). Predictive modeling of telecom customer churn using deep learning. *Journal of Machine Learning Research*, *24*, 1–15.

Karimi, S. (2021). Customer segmentation using clustering methods in telecom. *Telecommunications Review*, *29*(3), 45–53.

Kavitha, V. (2020). Role of machine learning in predicting customer churn in telecom industry. *International Journal of Advanced Computer Science and Applications*, *11*(6), 518–524.

Kingsakda, V., Duangvilaykeo, T., & Saybounheuangs, S. (2024). Factors affecting for using services of mobile phone network for customer of the Lao Telecommunications Company, Oudomxay Province. *Souphanouvong University Journal Multidisciplinary Research and Development, 10*(4), 56–64. https://doi.org/10.69692/SUJMRD100456

Kosaraju, N., Sankepally, S. R., & Mallikharjuna Rao, K. (2023, February). Churn prediction in telecom using ensemble methods. *Proceedings of the 7th International Conference on Data Science and Analytics* (pp. 55–60).

Lalwani, P., Mishra, S., & Chadha, D. (2021). Feature selection techniques for customer churn prediction: A comparative analysis. *Journal of Applied Data Science*, *2*(1), 44–53.

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., & Zhang, C. (2021). Towards automated AI model management in telecom data pipelines. *IEEE Transactions on Knowledge and Data Engineering*, *33*(11), 3249–3263.

Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, *55*(1), 1–39. https://doi.org/10.1145/3485128

Samuel, K. (1986). Modeling subscriber behavior in early telecom networks. *Telecom Systems Journal*, *12*(3), 201–218.

Saypadith, S., & Onoye, T. (2021). An edge-AI framework for real-time customer churn prediction. *IEEE Access*, *9*, 78865–78875. https://doi.org/10.1109/ACCESS.2021.3085076

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Responsible AI pipelines: Bias, explainability, and governance. *Proceedings of the VLDB Endowment*, *16*(9), 2251–2264.

Sruthi, E. R. (2024). Churn detection in the telecom sector using hybrid deep learning models. *International Journal of Data Science*, *5*(2), 110–121.

Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). Deep learning-based customer behavior prediction in telecom. *Journal of Information Technology Research*, *14*(4), 92–104.

Wagh, K., & Andhale, A. (2024). Comparative analysis of machine learning techniques for telecom churn prediction. *Journal of Emerging Technologies*, *12*(1), 25–31.

ตาตะลางที 1: ສະແດງໃຫ້ເຫັນເຖິງລາຍລະອຽດຂໍ້ມູນທີ່ມີ ຊື່ຕົວປ່ຽນ ແລະ ລາຍລະອຽດຂອງຕົວປ່ຽນ

ຊື່ຕົວປ່ຽນ	ລາຍລະອຽດຕົວປ່ຽນ
msisdn	ລະຫັດປະຈຳຕົວຂອງລູກຄ້າແຕ່ລະຄົນ
Refill	ຍອດເງິນທີ່ລູກຄ້າໄດ້ຕື່ມໃນຊ່ວງເວລານັ້ນ (Yes ຫຼື No)
Package	ຈຳນວນເງິນທີ່ໃຊ້ຈ່າຍໃນການຊື້ແພັກເກັດ (Yes ຫຼື No)
Call	ຍອດເງິນທີ່ໃຊ້ໃນການໂທລະສັບ (Yes ຫຼື No)
SMS	ຍອດເງິນທີ່ໃຊ້ໃນການສົ່ງຂໍ້ຄວາມ SMS (Yes ຫຼື No)
Vas	ຈຳນວນເງິນທີ່ໃຊ້ໃນບໍລິການສ່ວນເສີມ (VAS) (Yes ຫຼື No)
All_Amount	ຍອດເງິນທີ່ໃຊ້ລວມທັງໝົດໃນທຸກບໍລິການ
Tenure	ໄລຍະເວລາທີ່ລູກຄ້າໃຊ້ບໍລິການ (ເຊັ່ນ 1, 2, 3 ເດືອນ)
Gender	ເພດຂອງລູກຄ້າ
Customer Complain	ມີການຮ້ອງທຸກໃນການໃຊ້ບໍລິການ (Yes ຫຼື No)
Customer_Class	ການແບ່ງປະເພດລູກຄ້າຕາມລະດັບການໃຊ້ບໍລິການ ຫຼື ຄ່າໃຊ້ຈ່າຍ (ເຊັ່ນ "A", "B")
Churn	ຕົວປ່ຽນເປົ້າໝາຍ: ລູກຄ້າຢຸດໃຊ້ບໍລິການ (Yes) ຫຼື ຍັງໃຊ້ຢູ່ (No)

ตาตะลางที 2: ภานอัดผินภานปะเมินในภานคาดคะเนด้อย Confusion Matric

		ค่าความจิง	
		บ่ยุดใช้	ยุดใช้
ภานคาดคะเน	บ่ยุดใช้	TP	FP
	ยุดใช้	FN	TN

ตาตะลางที 3: ผินภานอัดความไอ และ ความแตกต่างในภานธวนธุ้ของเต้ากโนโลชีในภานธวนธุ้

Training Set (%)	ความไอ (s)	
	Logistic Regression	Random Forest
70	3.942	141.169
80	4.285	154.149
90	5.661	184.976

ตาตะลางที 4: ผินภานอัดแทกความไอ และ ความแตกต่างในภานทึดสอบของเต้ากโนโลชีในภานธวนธุ้

Testing Set (%)	ความไอ (s)	
	Logistic Regression	Random Forest
30	0.015	13.232
20	0.009	9.234
10	0.005	4.5888

ตาตะลางที 5: ผินภานปวบทวบค่าความถึกต้อง และ ความแตกต่างของเต้ากโนโลชีภานธวนธุ้

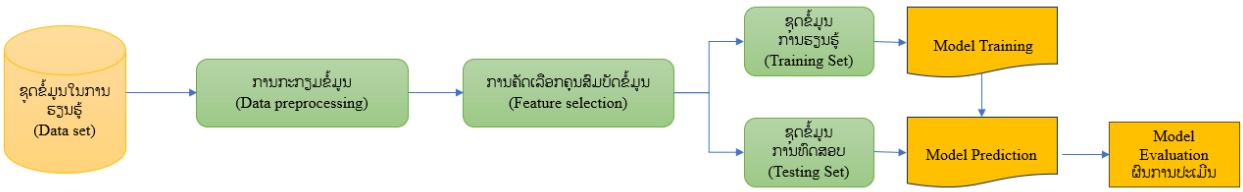
Training - Testing Set (%)	AUC	
	Logistic Regression	Random Forest
70-30	0.80	0.81
80-20	0.81	0.81
90-10	0.80	0.81

ตาตะลางที 6: ผินภานปะเมินปะสิดทีพาบ และ ความแตกต่างของเต้ากโนโลชีภานธวนธุ้

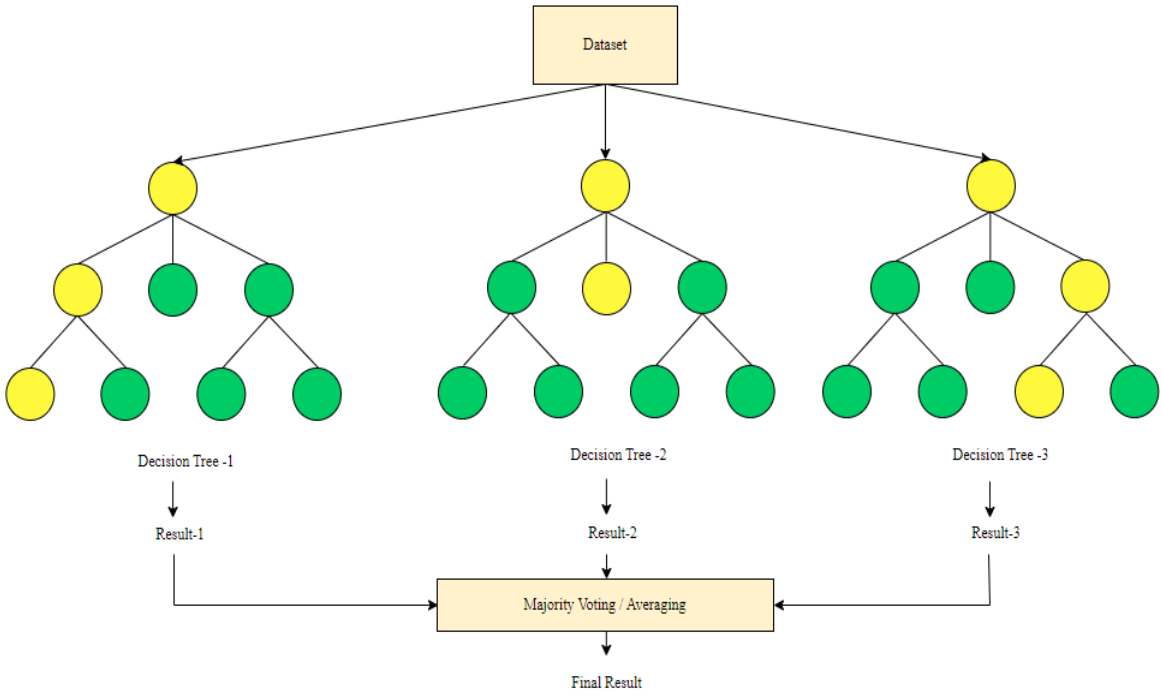
Training – Testing Set (%)	**Logistic Regression**				**Random Forest**			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
70-30	0.87	0.87	0.85	0.86	0.87	0.87	0.87	0.87
80-20	0.88	0.87	0.85	0.86	0.87	0.88	0.87	0.87
90-10	0.87	0.87	0.85	0.86	0.87	0.88	0.87	0.87

ຕາຕະລາງທີ 7: Confusion Matrix ຂອງຜົນການຄາດຄະເນການຢຸດໃຊ້ງານ ລະຫວ່າງ Logistic Regression and Random Forest

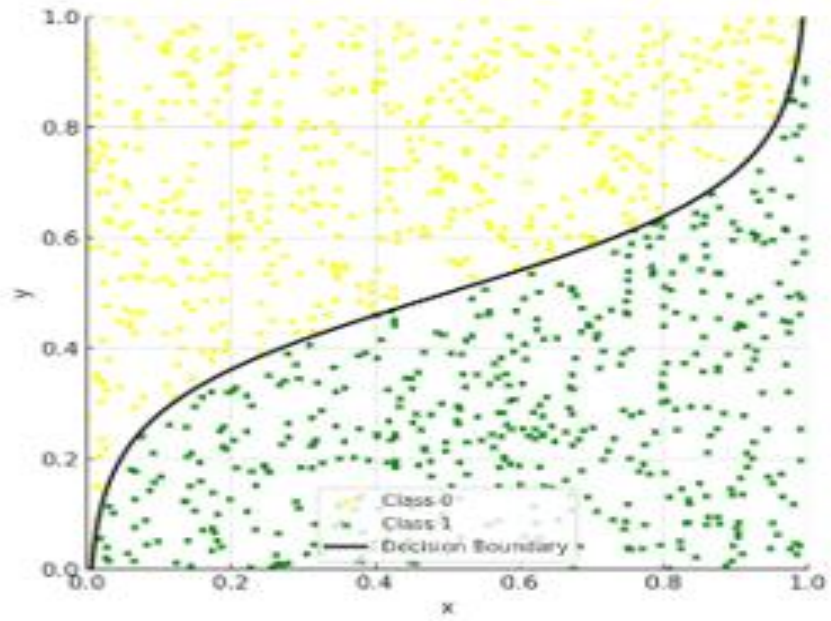
Training - Testing Set (%)	Logistic Regression		Random Forest	
70-30	233731	2022	227048	8705
	38273	26905	29040	36138
80-20	155817	1352	151730	5439
	25383	18069	19268	24184
90-10	77908	677	75961	2624
	12725	9001	9755	11971



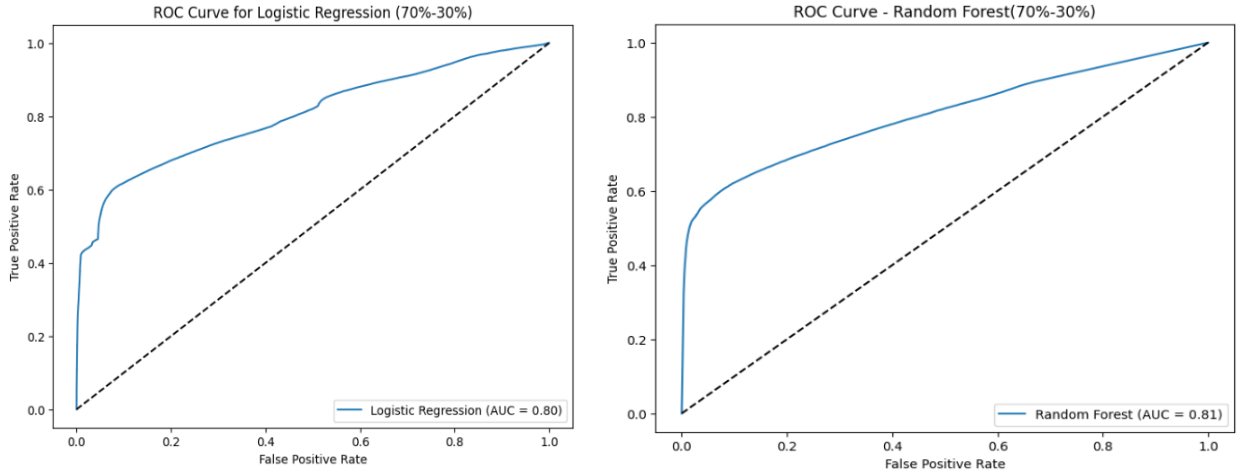
ຮູບທີ 1: ແຜນວາດຂັ້ນຕອນໃນການປະເມີນປະສິດທິພາບຂອງເຕັກໂນໂລຊີການຮຽນຮູ້



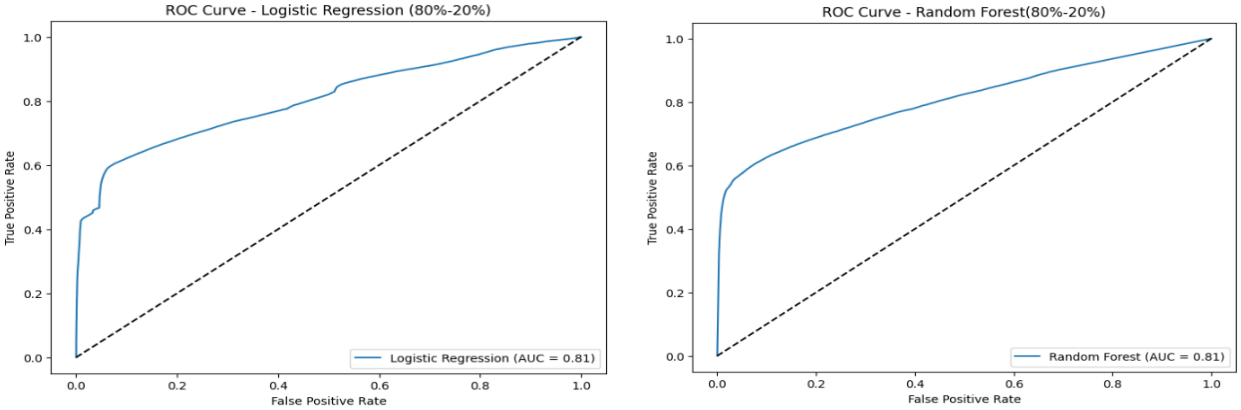
ຮູບທີ 2: ຮູບແບບການເຮັດວຽກຂອງ Random Forest



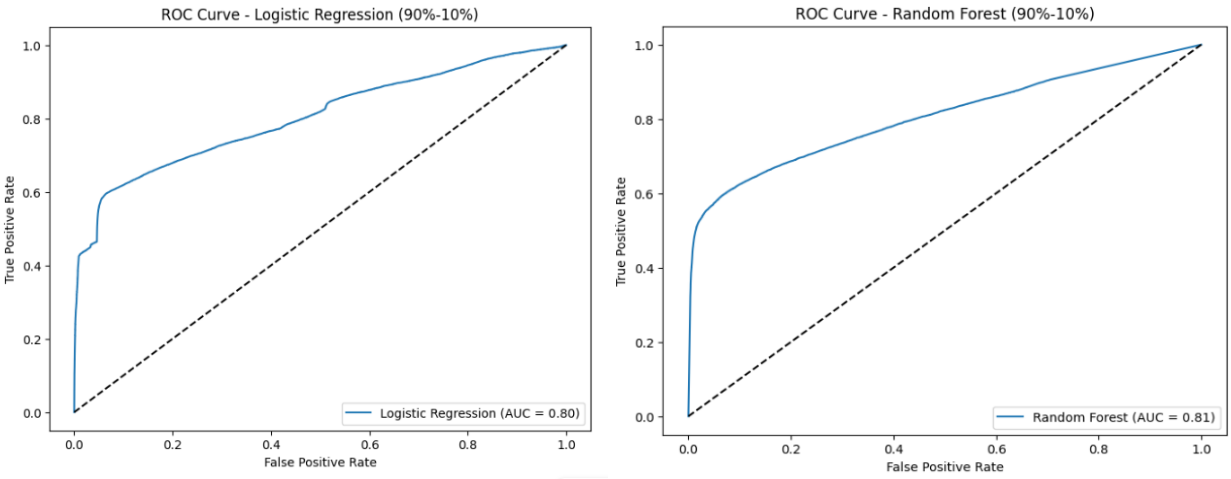
ຮູບທີ 3: ຮູບແບບການເຮັດວຽກຂອງ Logistic Regression ໂດຍໃຊ້ Sigmoid Function



ຮູບທີ 4: ແຜນວາດຄ່າຄວາມຖືກຕ້ອງໃນການຄາດຄະເນຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ (70%-30%)



ຮູບທີ 5: ແຜນວາດຄ່າຄວາມຖືກຕ້ອງໃນການຄາດຄະເນຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ (80%-20%)



ຮູບທີ 6: ແຜນວາດຄ່າຄວາມຖືກຕ້ອງໃນການຄາດຄະເນຂອງເຕັກໂນໂລຊີການຮຽນຮູ້ (90%-10%)