CSC 635 Data Mining

## **Project Proposal**

### Submitted to:

### Dr. Jamil Saquer

### Author:

### Aileen Bui

### Manik Vashith

### Rafail Islam

**DATASET ON THE APPOINTMENT NO-SHOW RATE AT 5 CLINICS IN DES MOINES METRO AREA, IOWA, USA**

1. **Introduction**

This dataset was obtained from a healthcare company, which contains the data of appointments scheduled in the Des Moines Metro area for 5 different clinics in 2018. The dataset was provided to the School of Business of Missouri State University in April 2019 for use of MBA students in doing a business case study. The dataset has the label is the status of the appointment, which is either show or no-show.

By using the dataset to build up a model to predict whether a patient will show up at the appointment or not, we believe that it will help improve the operational efficiency of the clinics where they can have a waitlist for appointment that likely will be cancelled. The goal of the healthcare company is to maximize its capability to serve patients. This dataset has 95,221 records and is a skewed dataset with only roughly 10% of records are the positive cases (patients do not show up). This typically represents real life data from healthcare industry where the positive class is the minority class.

1. **Dataset description**

The dataset has each row is an appointment with 33 features as shown in Table 1 in Appendix. We have accessed the dataset to identify data issues (such as negative value as data noise in the duration of appointment, missing values, etc.). The pre-processing of this data also includes choosing whether to do over-sampling or under-sampling for the dataset in order to handle the skewness in the data.

The original dataset is attached together with this proposal in the submission.

1. **Project plan**

In terms of algorithm, we prioritize using the algorithm that can do features selection and explain the results rather than a black box because in healthcare industry practice, similar to banking industry with credit score prediction, the ability for the model to explain its decision is as important as a high accuracy rate. By identifying the features that were used in the prediction, business leaders can customize different strategies to improve those features. Hence, we plan to apply decision tree, random forest and maybe some clustering algorithms to discover the hidden patterns in the data.

**APPENDIX**

Table 1: List of variables in the clinical appointments datase

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| AppointmentID | Unique ID for each appointment made. |
| PatientID | Unique ID for each patient. A patient may make multiple appointments. |
| ClinicNM | Name of Clinic (A to E) |
| AppointmentDTS | Date of Appointment |
| AppointmentMonthNBR | Month of Appointment |
| AppointmentWeekdayNBR | Weekday of Appointment (1 = Sunday, 7 = Saturday) |
| AppointmentHourNBR | Hour of Appointment |
| AgeNBR | Patient Age When Appointment Was Made |
| SexFLG | Sex of Patient |
| HispanicFLG | Whether or not patient is Hispanic (0 = no, 1 = yes) |
| SingleFLG | Whether or not patient lives alone (0 = no, 1 = yes) |
| LivesInApartmentFLG | If patient lives in Apartment (0 = no, 1 = yes) |
| EmailFLG | Whether or not patient provided an email (0 = no, 1 = yes) |
| ApptLagNBR | Number of days between date appointment made and actual appointment date |
| InsuranceDSC | Insurance of Patient |
| HypertensionFLG | Whether or not patient has a known history of hypertension (0 = no, 1 = yes) |
| AsthmaFLG | Whether or not patient has a known history of asthma (0 = no, 1 = yes) |
| HeartDiseaseFLG | Whether or not patient has a known history of heart disease (0 = no, 1 = yes) |
| ObeseFLG | Whether or not patient has a known history of obesity (0 = no, 1 = yes) |
| DiabetesFLG | Whether or not patient has a known history of obesity (0 = no, 1 = yes) |
| Noshow24NBR | Number of times patient no-showed in appointments in last 24 months prior to appointment made date |
| CancellationsNBR | Number of times patient cancelled appointments within 24 hours of appointment in last 24 months prior to appointment made date |
| Latearrivals24NBR | Number of times patient arrived late to appointments in last 24 months prior to appointment made date |
| CheckintoCheckoutNBR | Average number of minutes between patient check in to patient check out in completed appointments last 24 months prior to appointment made date |
| AppttoCheckoutNBR | Average number of minutes between original appointment time to actual patient check out in completed appointments last 24 months prior to appointment made date |
| CheckintoApptNBR | Average number of minutes between patient check in to original appointment time in completed appointments last 24 months prior to appointment made date |
| Arrived24NBR | Number of times patient arrive to an appointment in the 24 months prior to appointment made date |
| Providers24CNT | Number of distinct providers the patient has seen in the 24 months prior to appointment made date |
| ThatProvider24NBR | Number of times the patient has had an appointment with this provider in the 24 months prior to appointment made date |
| NoshowRate24NBR | The percentage of appointments the patient has no-showed in the 24 months prior to appointment made date |
| EdVisitsNBR | Number of Emergency Department visits the patient has had in the 12 months prior appointment made date |
| IpVisitsNBR | Number of Inpatient Admissions the patient has had in the 12 months prior appointment made date |
| NoShowFLG | If the patient no-showed for this appointment (0 = no, 1 = yes) |
| CancelledLateFLG | If the patient cancelled within 24 hours of this appointment (0 = no, 1 = yes) |