PHASE 2 TEXT OF VIDEO PRESENTATION

SPEAKER: PRAGAT WAGLE
SLIDE 1-3
HELLO WE ARE GROUP 10 AND WE CHOSE TO DO OUR PROJECT ON HOME CREDIT DEFAULT RISK PREDICTION. MY NAME IS PRAGAT AND I AND I WORKED WITH SAHIL, MARK, KUNAL ON THIS PROJECT. HOME CREDIT SET UP A COMPETITION ON KAGGLE TO USE MACHINE LEARNING TO ENSURE CLIENTS CAPABLE OF LOAN REPAYMENT ARE NOT REJECTED LOANS. THIS PROJECT AIMS TO IMPROVE FINANCIAL INCLUSION BY INCLUDING TELCO AND TRANSACTIONAL DATA TO PREDICT CLIENT REPAYMENT ABLITY. THIS IS AN EXAMPLE OF ONE OF VISUAL EDA'S WE SUED WHICH CAPTURES, CORRELATION WHICH IS A STATISTICAL MEASURE THAT EXPRESSES THE EXTENT TO WHICH TWO VARIABLES ARE LINEARLY RELATED (MEANING THEY CHANGE TOGETHER AT A CONSTANT RATE)

SPEAKER: MARK GREEN
SLIDE 4-5
HI MY NAME IS MARK AND BELOW WE HAVE SOME EXAMPLES OF VISUAL EXPLORATORY DATA ANALYSIS USING HISTOGRAMS, BOXPLOTS, AND KERNEL DENSITY ESTIMATES TO EXPLORE THE DISTRIBUTIONS OF SOME OF THE VARIABLES IN OUR DATASETS. THE INSIGHTS FROM THESE EDAS INFORM DECISIONS IN OUR MODELING PIPELINE. BELOW YOU CAN SEE THE BASELINE MODEL IMPUTING DATA WHERE THERE ARE MISSING DATA, SCALING NUMERICAL VARIABLES WHERE LARGE VALUES CAUSE PROBLEMS, AND ONE-HOT-ENCODING CATEGORICAL VARIABLES.

SPEAKER: SAHIL DHINGRA
SLIDE 6-7
WITH SIMPLE PIPELINE, OUR BASELINE SCORE FOR PHASE 1 WAS 74.34% AND KAGGLE SCORE WAS 73.3. WE CREATED SOME NEW FEATURES AND DID LOW LEVEL ANALYSIS BY DOING EDA AND PLOTTING. ROLLING UP ALL THE DATA TO TEST AND TRAIN, RESULTED IN KAGGLE SCORE OF 76.23. FOCUS AREA FOR THIS PHASE WAS EDA, ROLLING UP AGGREGATING DATA AND DERIVING INSIGHTS FROM STATISTICAL PLOTS. THIS HELPED US DRASTICALLY IMPROVE ROC SCORE. IN NEXT WE ARE PLANNING TO DEEP DIVE A LITTLE MORE IN FEATURE ENGINEERING AND HYPER PARAMETER TUNING TO FURTHER IMPROVE OUR SCORE.

SPEAKER: KUNAL SINGH
SLIDE 8
PHASE 1 WAS MOSTLY PLANNING AND EVALUATION BASED UPON DOMAIN ANALYSIS. WITH BASIC UNDERSTANDING, WE RAN OUR 1ST LOGISTIC REGRESSION MODEL. IN THIS PHASE, WE DID FEATURE ENGINEERING AND PLOTTING. ROLLING UP AGGREGATED DATA HELPED US IMPROVE THE ROC SCORE. IN PHASE 3, WE ARE PLANNING TO FURTHER UP-THE-ANTE ON FEATURE ENGINEERING AND INCLUDE HYPER-PARAMETER TUNING TO GET THE BEST

FEATURES AND PARAMETERS. MAJOR CONCERN WAS SHEER DATA SIZE, WE WERE AGGREGATING ON SEVEN PARAMETERS WHICH MEANT NOW OUR DATA WAS 7X NOT COUNTING THE OHE COLUMNS X7. WE NEED TO SEE HOW WE CAN MANAGE RESOURCES OR USE PAID GOOGLE COLLAB FEATURES.