Write a brief summary on these topics related to this problem:  
  
**1. Feature engineering:**  
○ How do you propose to turn the data into something we can do predictive modeling?

I have used TFIDF to create a model. But I think better solution would be using Transfer Learning. But due to shortage of time for the assignment, I was only able to try traditional models.

○ Deliverables: features list and calculations, packages used

Feature list is present in the feature list file in the folder. Packages used are: -

1. sklearn (TfidfVectorizer, PCA, LabelEncoding)

Feature list : ['000', '000 epinephrine', '01', '02', '03', '04', '05', '06', '07', '08', '09', '10', '100', '100 000', '100 000 epinephrine', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '200', '2007', '2008', '2009', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '325', '40', '60', '6h', '70', '75', '92', '93', '95', '96', '97', 'abc', 'abc patient', 'abdomen', 'abdomen abdomen', 'abdomen bowel', 'abdomen bowel extremities', 'abdomen extremities', 'abdomen extremities edema', 'abdomen liver', 'abdomen patient', 'abdomen pelvis', 'abdomen skin', 'abdominal', 'abdominal patient', 'abdominal wall', 'accommodation', 'acetabulum', 'acid', 'acute', 'adenocarcinoma', 'adenoid', 'adenoma', 'adnexal', 'adrenal', 'adrenal gland', 'airway', 'albumin', 'albuterol', 'alcohol', 'alcohol patient', 'alkaline', 'allograft', 'amniotic', 'amniotic fluid', 'amoxicillin', 'anal', 'anesthesia', 'aneurysm', 'annular', 'anterior', 'anterior abdominal', 'anterior abdominal wall', 'anterior cervical', 'anterior chamber', 'anterior posterior', 'anteriorly', 'antrum', 'aorta', 'aortic', 'aortic arch', 'aortic valve', 'apical', 'appearing', 'arch', 'area', 'arm', 'arterial', 'artery', 'artery artery', 'artery carotid', 'artery carotid artery', 'artery circumflex', 'artery coronary', 'artery coronary artery', 'artery patient', 'artery pulmonary', 'artery pulmonary artery', 'artery vessel', 'articular', 'aspirin', 'ast', 'atenolol', 'atrial', 'atrium', 'axilla', 'axillary', 'b12', 'bacitracin', 'bag', 'basal', 'base', 'bed', 'bed patient', 'benign', 'beta', 'betadine', 'bicarbonate', 'biceps', 'bilateral', 'bilaterally', 'bile', 'biliary', 'bilirubin', 'biopsies', 'biopsy', 'biopsy patient', 'bladder', 'bladder bladder', 'bladder neck', 'bladder patient', 'bladder tumor', 'bleed', 'blocker', 'blood', 'blood blood', 'blood cell', 'blood heart', 'blood patient', 'blood patient patient', 'blood stool', 'blood sugar', 'blood urine', 'bloody', 'blunt', 'bluntly', 'bmp', 'body', 'bone', 'bone bone', 'bone graft', 'bone marrow', 'bone patient', 'bone wound', 'border', 'bowel', 'bowel bladder', 'bowel bowel', 'bowel extremities', 'bowel extremities edema', 'bowel patient', 'boy', 'brachial', 'brain', 'brain tumor', 'breast', 'breast cancer', 'breasts', 'bronchoscopy', 'bronchus', 'brother', 'bruits', 'bun', 'bursa', 'c3', 'c4', 'c4 c5', 'c4 c5 c5', 'c5', 'c5 c5', 'c5 c5 c6', 'c5 c6', 'c5 c6 c6', 'c6', 'c6 c6', 'c6 c6 c7', 'c6 c7', 'c7', 'calcium', 'calf', 'caliber', 'caliber vessel', 'canal', 'cancer', 'cancer patient', 'cannula', 'capillary', 'capsular', 'capsular bag', 'capsule', 'carcinoma', 'cardiac', 'cardiac coronary', 'cardiac patient', 'cardioplegia', 'cardiopulmonary', 'cardiovascular', 'cardiovascular heart', 'cardiovascular regular', 'cardiovascular regular abdomen', 'carotid', 'carotid artery', 'carotid artery carotid', 'carotid bruits', 'carpal', 'cartilage', 'cat', 'catheter', 'caudal', 'cava', 'cavity', 'cbc', 'cecum', 'cell', 'cell carcinoma', 'cell hemoglobin', 'central', 'cephalic', 'cerebellar', 'cerebral', 'cervical', 'cervical c5', 'cervical c5 c6', 'cervical cervical', 'cervical patient', 'cervical spine', 'cervix', 'chamber', 'chest', 'chest patient', 'chest tube', 'chest wall', 'children', 'chloride', 'cholesterol', 'circumferential', 'circumflex', 'circumflex artery', 'circumflex coronary', 'circumflex coronary artery', 'clot', 'cm', 'cns', 'codeine', 'colace', 'collateral', 'colon', 'colon cancer', 'colonic', 'conjunctiva', 'cord', 'cord cord', 'cornea', 'corneal', 'coronary', 'coronary artery', 'coronary artery coronary', 'coronary artery patient', 'coronary artery vessel', 'coronary coronary', 'cortex', 'cortical', 'coumadin', 'coumadin patient', 'count', 'cranial', 'cranial nerve', 'crease', 'creatinine', 'creatinine glucose', 'cricoid', 'crp', 'cruciate', 'csf', 'cuff', 'cul', 'cul sac', 'cyst', 'cystic', 'cystic duct', 'daily', 'day', 'dd', 'dd yyyy', 'decadron', 'deltoid', 'descending', 'digoxin', 'disease', 'dose', 'dr', 'dr patient', 'duct', 'duodenum', 'dura', 'ear', 'ear canal', 'ear ear', 'ear patient', 'ears', 'edema', 'edema patient', 'edge', 'endometrial', 'endotracheal', 'endotracheal patient', 'endotracheal tube', 'endplate', 'epidural', 'epinephrine', 'erythema', 'esophageal', 'esophagitis', 'esophagus', 'esophagus stomach', 'eustachian', 'extensor', 'extensor tendon', 'extra', 'extraocular', 'extraocular muscle', 'extremities', 'extremities edema', 'extremities patient', 'extremities skin', 'extremity', 'exudate', 'eye', 'eye eye', 'eye eye eye', 'eye eye patient', 'eye patient', 'eyelid', 'eyes', 'facial', 'fallopian', 'fallopian tube', 'fascia', 'fascia skin', 'fascial', 'fat', 'female', 'femoral', 'femoral artery', 'femur', 'fentanyl', 'fetal', 'fetal heart', 'fhx', 'fibroid', 'figure', 'flank', 'flap', 'flexor', 'fluid', 'focal', 'folic', 'folic acid', 'foot', 'foot foot', 'foramen', 'foreskin', 'fossa', 'free', 'french', 'french sheath', 'frontal', 'fundus', 'gadolinium', 'gag', 'gallbladder', 'gallbladder gallbladder', 'ganglion', 'gastric', 'gastritis', 'gastroesophageal', 'gastrointestinal', 'gauge', 'gentamicin', 'gentleman', 'gerd', 'gland', 'glans', 'glenoid', 'glucose', 'grade', 'graft', 'graft graft', 'guidewire', 'gut', 'hair', 'hand', 'hct', 'head', 'head neck', 'heart', 'heart abdomen', 'heart abdomen bowel', 'heart abdomen extremities', 'heart blood', 'heart coronary', 'heart heart', 'heart lungs', 'heart lungs abdomen', 'heart patient', 'hematoma', 'hemoglobin', 'hemoglobin platelet', 'heparin', 'hepatic', 'hepatitis', 'hepatosplenomegaly', 'hiatal', 'high', 'high grade', 'humerus', 'hydrochlorothiazide', 'hypertension', 'ibuprofen', 'ii', 'ileum', 'iliac', 'iliac artery', 'incised', 'incision', 'infant', 'inferior', 'inguinal', 'injected', 'inner', 'inr', 'insufficiency', 'insulin', 'intercostal', 'intraarticular', 'intracranial', 'intraocular', 'intrauterine', 'intravenous', 'intravenously', 'iodine', 'iris', 'iron', 'iv', 'jaw', 'joint', 'joint joint', 'joint patient', 'joint skin', 'judkins', 'jugular', 'jugular vein', 'jugular venous', 'kerrison', 'kidney', 'kidney kidney', 'kidney patient', 'knee', 'knee patient', 'l3', 'l4', 'l4 l5', 'l4 l5 l5', 'l5', 'l5 l5', 'l5 l5 s1', 'l5 s1', 'labrum', 'lady', 'lamina', 'large', 'laryngeal', 'larynx', 'lasix', 'lateral', 'lavage', 'layer', 'left', 'left ventricular', 'leg', 'leg patient', 'lens', 'lesion', 'lesions', 'levaquin', 'lidocaine', 'lidocaine 100', 'lidocaine 100 000', 'lidocaine epinephrine', 'ligament', 'light', 'light accommodation', 'light extraocular', 'limb', 'limbus', 'line', 'lipitor', 'lisinopril', 'liver', 'liver patient', 'liver spleen', 'lobe', 'low', 'low grade', 'lumbar', 'lumbosacral', 'lumen', 'lung', 'lung cancer', 'lung patient', 'lungs', 'lungs abdomen', 'lungs abdomen bowel', 'lungs cardiac', 'lungs heart', 'lungs heart abdomen', 'lymph', 'lymph node', 'lymphadenopathy', 'lymphatic', 'lymphocyte', 'lymphoma', 'magnesium', 'mammary', 'mammary artery', 'man', 'mandible', 'marcaine', 'marcaine epinephrine', 'marcaine patient', 'marrow', 'material', 'maxillary', 'meatus', 'medial', 'medial lateral', 'mediastinal', 'mediastinum', 'medrol', 'membrane', 'meniscus', 'mesothelioma', 'metatarsal', 'metatarsal head', 'metatarsophalangeal', 'metatarsophalangeal joint', 'metformin', 'methotrexate', 'metoprolol', 'mg', 'midas', 'midline', 'mitral', 'mitral valve', 'mm', 'mm dd', 'mm dd yyyy', 'morphine', 'mouth', 'mucosa', 'mucous', 'mucous membrane', 'multivitamin', 'muscle', 'muscle fascia', 'muscle muscle', 'muscle patient', 'muscle skin', 'muscular', 'musculature', 'myocardial', 'nail', 'nares', 'nasal', 'nasal mucosa', 'nasal septum', 'nasopharynx', 'neck', 'needle', 'needle patient', 'neoplasm', 'nerve', 'nerve patient', 'nerve root', 'neuroma', 'neurovascular', 'nipple', 'nitroglycerin', 'node', 'nodule', 'non', 'normal', 'norvasc', 'nose', 'notch', 'nuclear', 'nucleus', 'oblique', 'obtuse', 'oil', 'ointment', 'old', 'old gentleman', 'old patient', 'omentum', 'optic', 'oral', 'oral mucosa', 'oral patient', 'organ', 'organomegaly', 'oropharynx', 'osteophyte', 'ostium', 'ovarian', 'ovary', 'oxygen', 'oxygen patient', 'pad', 'pain', 'palate', 'pancreas', 'pancreatic', 'pap', 'papillary', 'parathyroid', 'paresthesia', 'parietal', 'patch', 'patella', 'patellar', 'patient abc', 'patient abdomen', 'patient abdominal', 'patient alcohol', 'patient alcohol patient', 'patient anterior', 'patient bed', 'patient betadine', 'patient biopsy', 'patient bladder', 'patient blood', 'patient blood patient', 'patient bone', 'patient bowel', 'patient brain', 'patient breast', 'patient cardiac', 'patient cervical', 'patient chest', 'patient coronary', 'patient coronary artery', 'patient coumadin', 'patient dr', 'patient ear', 'patient endotracheal', 'patient eye', 'patient eyes', 'patient female', 'patient gentleman', 'patient head', 'patient heart', 'patient hemoglobin', 'patient intravenous', 'patient joint', 'patient kidney', 'patient knee', 'patient leg', 'patient lung', 'patient marcaine', 'patient muscle', 'patient nasal', 'patient oral', 'patient oxygen', 'patient patient', 'patient patient abdomen', 'patient patient alcohol', 'patient patient blood', 'patient patient cardiac', 'patient patient patient', 'patient patient skin', 'patient pulmonary', 'patient renal', 'patient scar', 'patient skin', 'patient spinal', 'patient thyroid', 'patient tobacco', 'patient tube', 'patient tylenol', 'patient urine', 'patient vaginal', 'patient vascular', 'patient wound', 'pectoralis', 'pedicle', 'pelvic', 'pelvis', 'penicillin', 'penis', 'pericardial', 'pericardium', 'perineum', 'periosteal', 'periosteum', 'peripheral', 'peripheral vascular', 'peritoneal', 'peritoneal cavity', 'peritoneum', 'pharynx', 'phenergan', 'pillar', 'pituitary', 'placenta', 'plantar', 'platelet', 'platelet count', 'platysma', 'plavix', 'pleura', 'pleural', 'po', 'polyp', 'portal', 'posterior', 'potassium', 'potassium chloride', 'pouch', 'prednisone', 'prepped', 'pronator', 'prostate', 'prostate cancer', 'protonix', 'ptt', 'pulmonary', 'pulmonary artery', 'pulmonary artery pulmonary', 'pulmonary patient', 'pulmonary pulmonary', 'pylorus', 'quadrant', 'radial', 'reactive', 'reactive light', 'reactive light accommodation', 'reactive light extraocular', 'reapproximated', 'rectal', 'rectum', 'rectus', 'red', 'regular', 'regular abdomen', 'renal', 'renal artery', 'renal patient', 'renal renal', 'retinal', 'retractor', 'retroperitoneal', 'rib', 'right', 'right coronary', 'root', 's1', 'sac', 'saline', 'saphenous', 'saphenous vein', 'saphenous vein graft', 'scalp', 'scar', 'scar tissue', 'scarpa', 'scleral', 'scrotal', 'scrotum', 'seed', 'septal', 'septum', 'serous', 'serum', 'shaft', 'sheath', 'shortness', 'sigmoid', 'sigmoid colon', 'silk', 'sinus', 'skin', 'skin blood', 'skin edge', 'skin muscle', 'skin patient', 'skin patient patient', 'skin skin', 'skin subcutaneous', 'skin subcutaneous tissue', 'skin wound', 'skull', 'small', 'smooth', 'sodium', 'soft', 'soft tissue', 'solid', 'space', 'specimens', 'spermatic', 'spinal', 'spinal cord', 'spinal needle', 'spine', 'spine cervical', 'spine patient', 'spinous', 'spleen', 'sputum', 'squamous', 'squamous cell', 'squamous cell carcinoma', 'staple', 'stem', 'steri', 'steri strips', 'sternocleidomastoid', 'sternocleidomastoid muscle', 'sternum', 'steroid', 'stomach', 'stone', 'stool', 'stricture', 'strips', 'subacromial', 'subclavian', 'subclavian vein', 'subcutaneous', 'subcutaneous fat', 'subcutaneous tissue', 'subcutaneous tissue skin', 'subcuticular', 'subperiosteal', 'sugar', 'sugar blood', 'superficial', 'supraclavicular', 'suprapubic', 'surface', 'suture', 'sweat', 'syndrome', 'synthroid', 't11', 'tablet', 'tear', 'teeth', 'teeth teeth', 'temporalis', 'tender', 'tenderness', 'tendon', 'tendon tendon', 'testicle', 'testicular', 'testis', 'thecal', 'thecal sac', 'thigh', 'thoracic', 'thoracic spine', 'throat', 'thrombus', 'thrush', 'thyroid', 'thyroid gland', 'tibia', 'tibial', 'tissue', 'tissue patient', 'tissue skin', 'tissue tissue', 'tissue wound', 'tobacco', 'tobacco alcohol', 'tongue', 'tonsil', 'tonsillar', 'tooth', 'toprol', 'tourniquet', 'trachea', 'tracheal', 'tract', 'tricuspid', 'trocar', 'troponin', 'tsh', 'tube', 'tube ovary', 'tube patient', 'tube tube', 'tuberosity', 'tumor', 'tumor patient', 'tumor tumor', 'turbinate', 'tylenol', 'tympanic', 'tympanic membrane', 'type', 'uihc', 'ulcer', 'ulnar', 'ulnar nerve', 'umbilical', 'umbilicus', 'upper', 'ureter', 'ureteral', 'urethra', 'urethral', 'urinary', 'urinary tract', 'urine', 'urine patient', 'uterine', 'uterus', 'vagina', 'vaginal', 'vaginal patient', 'valve', 'vancomycin', 'vascular', 'vascular patient', 'vein', 'vein graft', 'vena', 'vena cava', 'venous', 'ventral', 'ventricle', 'ventricular', 'ventricular patient', 'vertebral', 'vertebral body', 'vessel', 'vessel coronary', 'vicodin', 'vicryl', 'vicryl suture', 'vitamin', 'volar', 'wall', 'water', 'wbc', 'woman', 'wound', 'wound patient', 'wound saline', 'wound skin', 'wound wound', 'wound wound wound', 'xylocaine', 'year', 'year old', 'year old gentleman', 'year old patient', 'yyyy']

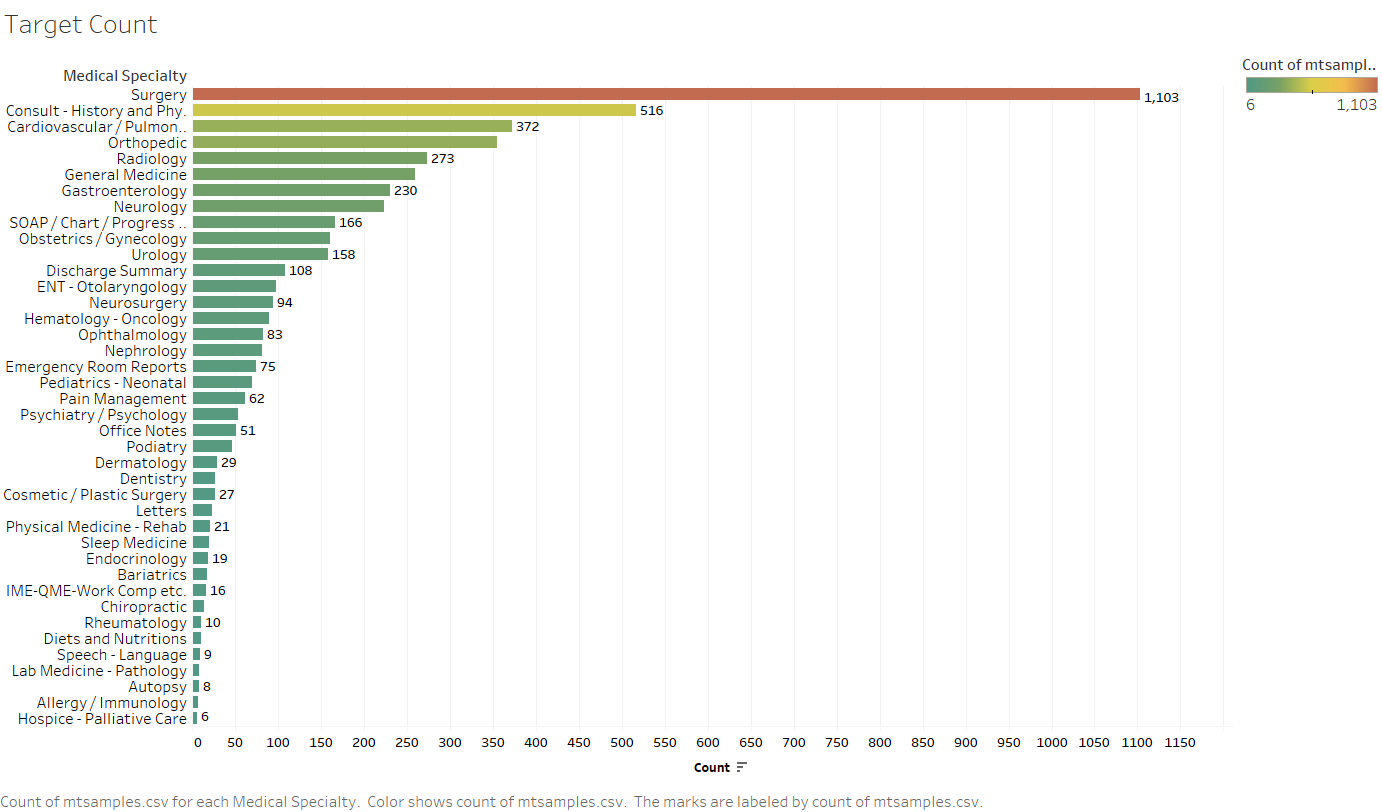
**2. Handling missing values:**   
○ We may design features based on their availability across samples. What happens when new data comes without these features?

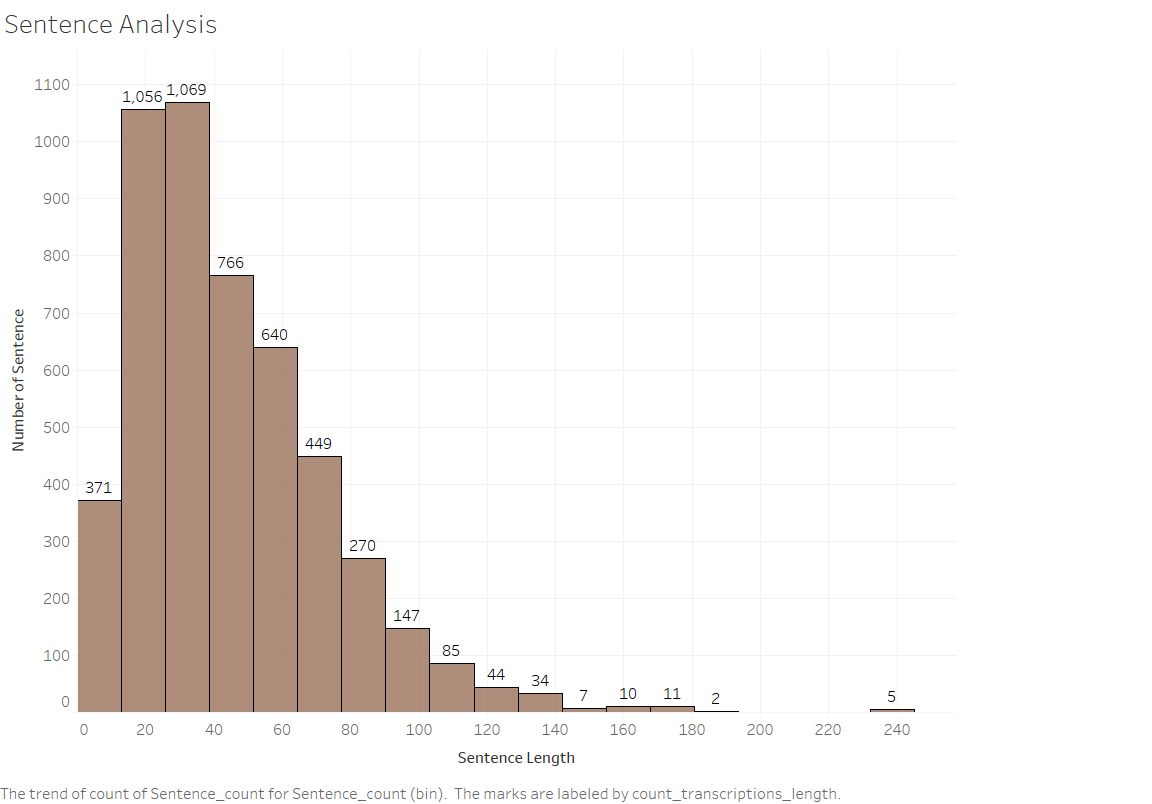
Same object on which we have used fit method on training data can be used. As that object have transform attribute which can be applied multiple time to the new data. This object internally create a vocabulary which is used to match with the new data and operations are hence done taking into that account.  
○ Deliverables: Notes, list methods with descriptions, packages used

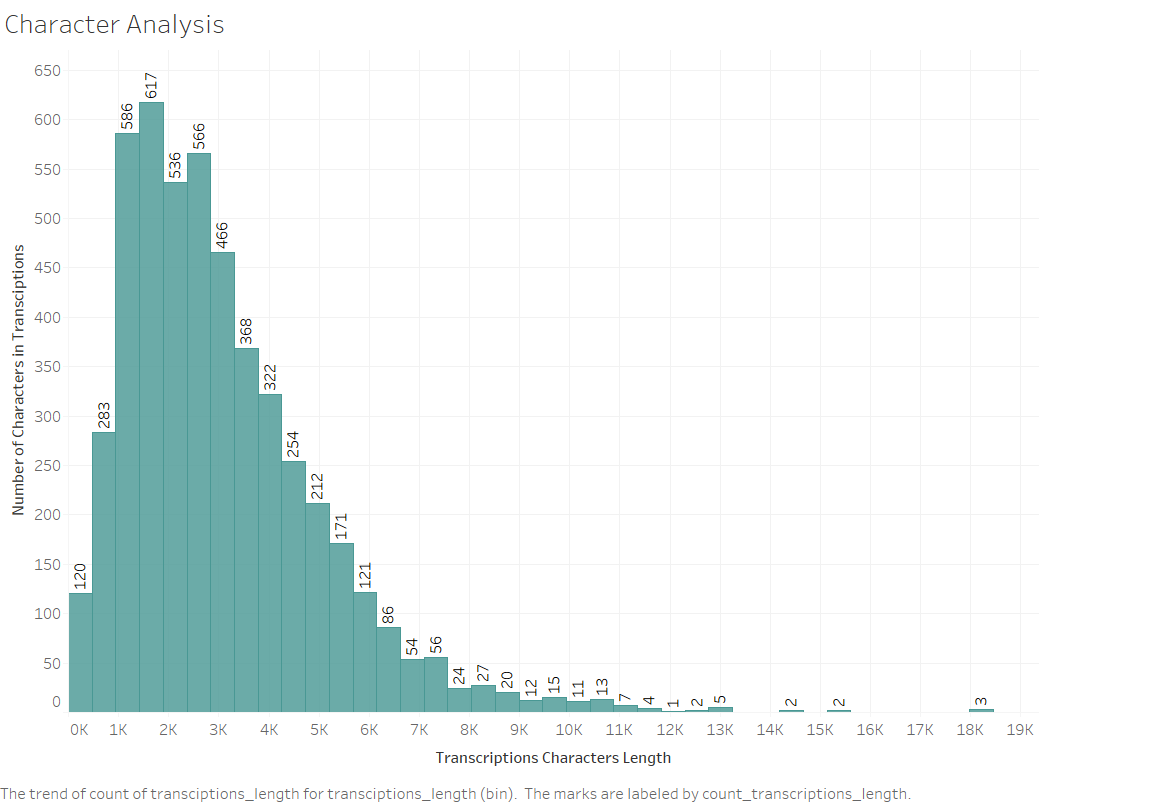
I have used sciscpacy models to detect medical entities in this text for processing biomedical or clinical text. As text has only 33 nan values present in transcription, I have deleted those records.

Packages used: -

1. Nltk (Lemmatizer, stopwords)
2. String (punctuation)
3. Pandas







**3. Handling the data imbalance in the data:**  
○ Some targets may have very few samples while some others will have a lot of samples,

In the given data, many classes have less than 50 records. For the improvement of the model these classes can be dropped from the dataset to train the model on the classes with majority records. But as it is a client decision. I have kept those classes to create the model. (This will have negative impact on scoring of the model.)  
○ Deliverables: methods and descriptions with block diagrams, packages

Some of the commonly used methods are under sampling, oversampling and SMOTE. As SMOTE is better techniques among these, I have used smote to balance the data.

Packages used are: -

1. Imblearn

**4. Modeling pipeline:**   
○ Splitting train, val and test data and training models and fine tuning the hyperparameters   
○ Deliverables: flow charts, model names and hyper parameters

I have used Optuna library for hyperparameter tuning. As it uses Bayesian optimization and tree parzen algorithm which are very fast to tune the parameters.

Model name: - Catboost (It does not need hyperparmeter tuning), XGBoost

XGBoost model

Number of finished trials: 100

Best trial:

F1\_score Value: 0.46494708994708994 (After dropping minority classes value comes out to be more than 60)

Params:

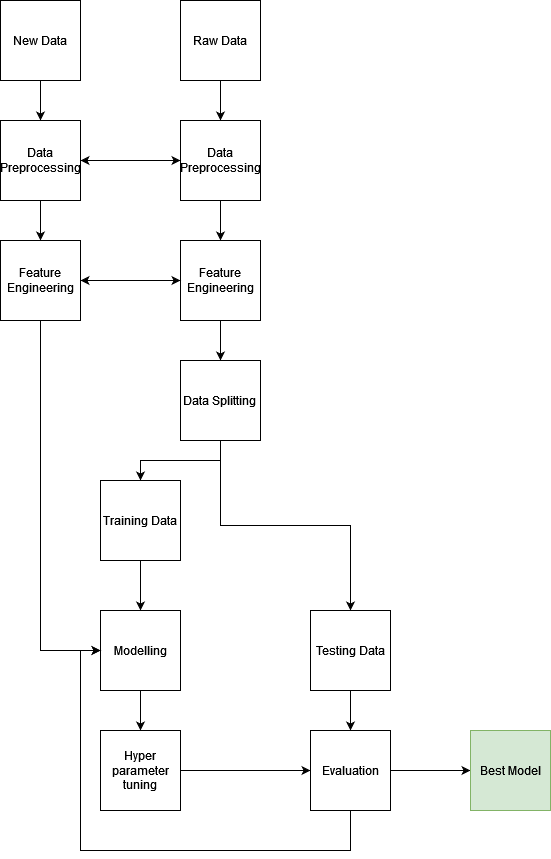
n\_estimators: 5715

booster: gblinear

lambda: 2.7256031470879e-06

alpha: 0.0003585174992460033

**5. MLOps:**   
○ Skim through this MLOps blog and write a brief proposal on an architecture for our problem which would ingest new data periodically and deploy new models when required   
○ Deliverables: descriptions and Flow chart and packages

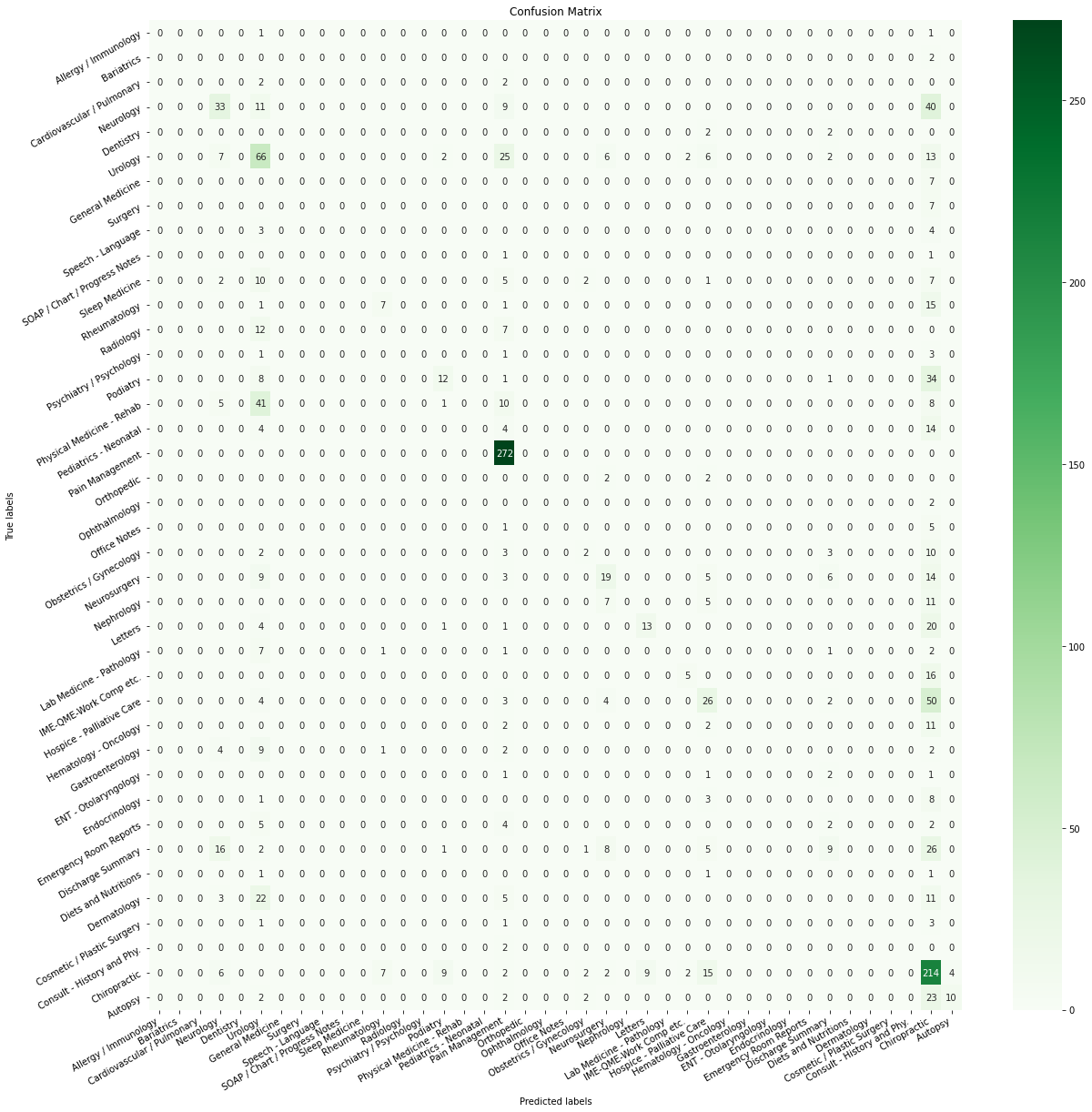
  
  
**6. Identifying and handling bias:**   
○ like gender, age or race related bias   
○ Deliverables: Notes

Bias is removed by using data imbalanced techniques as target variable has minority and majority classes. To remove bias, I have used SMOTE.  
 **7. What special features do you identify for this medical text data:**   
○ may be look at the vocabulary, abbreviations etc.   
○ Deliverables: Notes

Additionally, to improve the model we can use AbbreviationDetector of scispacy which is trained on biomedical text. We can also used gender details in the text as some diseases are gender specific. We can also parse age with regex.

Note: - Analysis on data was done in Tableau, link for the dashboard is: -

<https://public.tableau.com/app/profile/yogesh.kumar.singla/viz/Book2_16437355511910/Dashboard1>



**Conclusion**: Some categories are predicted well but minority classes are not predicted well. After dropping minority classes, results are going to improve signification. I have commented the code to drop minority classes but it can be uncommented and threshold can be provided according to the business needs.