DSC 680 Applied Data Science (2237-1)
Bellevue University
Professor Williams
Milestone 1 – Portfolio
Jake Meyer
07/02/2023

10 Portfolio Projects

The following list outlines the top projects or assignments I've chosen to showcase in my personal portfolio. These were completed during the M.S. Data Science Program offered by Bellevue University. In the future, I hope to add some personal projects to this portfolio.

1. Sales Forecasting and Late Delivery Prediction with Supply Chain Data
    a. Origin: Project 1 from DSC 680 Applied Data Science Course
    b. Description: The focus for this project was centered around Supply Chain Analytics. The two main focuses were Sales Forecasting and Late Delivery Prediction with data from DataCo from Mendeley Data. The CRISP-DM methodology was followed with Data Understanding, Data Preparation, Predictive Modeling, Evaluation, and Deployment stages. The Sales Forecasting problem was treated as a regression issue and the Late Delivery Prediction was treated as a classification problem. Models used for Sales Forecasting included Linear Regression, Lasso Regression, Decision Tree Regressor, and Random Forest Regressor. Models used for Late Delivery Prediction included Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. Metrics were utilized to compare each respective model and recommend a deployment decision/plan.
2. Prediction of Water Quality
    a. Origin: Project from DSC 630 Predictive Analytics
    b. Description: The focus for this project was to construct a model to predict whether water was potable based on certain water quality measurements. The data set used for this analysis was from Kaggle (Aditya Kadiwal). The CRISP-DM methodology was followed with Data Understanding, Data Preparation, Predictive Modeling, Evaluation, and Deployment stages. The classification problem had five potential models trained and evaluated: Support Vector Machine, K-Nearest Neighbor, Random Forest Classifier, Adaboost, Decision Tree Classifier, and Logistic Regression. The models were compared based on specific evaluation metrics and recommendations/insights drawn from the analysis.
3. Predicting Fuel Efficiency
    a. Origin: Assignment from DSC 550 Data Mining (Week 4)
    b. Description: This assignment focused on building a Linear Regression model to predict fuel efficiency (miles per gallon) of automobiles. The auto-mpg dataset from Kaggle (UCI Machine Learning) was utilized. The high-level steps within the assignment were prepping the data for the model, creating a correlation coefficient matrix, performing some exploratory data analysis, and training and testing a Linear Regression model. A Lasso Regression model was also trained and tested to compare with evaluation metrics such as R^2, RMSE, and MAE.
4. Sentiment Analysis Model with Movie Reviews
    a. Origin: Assignment from DSC 550 Data Mining (Week 5)
    b. Description: This assignment focused on building a Logistic Regression model for prediction of sentiment from movie reviews. The data set used was from 50,000 records of the IMDB movie reviews. The high-level steps within the assignment were getting stemmed data to prepare for training/testing the model, fitting and applying tf-df

vectorization to the training data, applying tf-df vectorization to the test data, training/testing a logistic regression model, determining metrics (accuracy, recall, precision, F1-score), plotting a confusion matrix, creating an ROC curve, and lastly repeating the training/test steps with a Multinomial Naïve Bayes model.

5. House Price Prediction and Mushroom Classification with Dimensionality Reduction and Feature Selection
    a. Origin: Assignment from DSC 550 Data Mining (Week 6)
    b. Description: There were two main objectives for this assignment. The idea was to show how a model performs when the most important features were included compared to all features. First, used Principal Component Analysis and Variance Threshold with a regression problem. House Prices data set from Kaggle was used for this portion of the assignment. The high-level steps for the regression problem were performing some data wrangling steps, identifying numerical vs. categorical features, preparing the data for training the model, application of feature selection methods of Principal Component Analysis and Feature Selection, and compared the model performance with R^2 and RMSE evaluation metrics. The second main objective was to apply the feature reduction techniques with a classification problem. The data set used for this portion of the assignment was from Kaggle (UCI Machine Learning) and focused on mushroom classification. This portion of the assignment consisted of converting categorical features to dummy variables, preparing the data, fitting a Decision Tree classifier on the training set, reporting the accuracy of the model, showing a visualization of the model, and using a Chai-statistic selector to identify the five best features for the model and repeating the accuracy evaluation of the model with only those features.
6. Loan Status Prediction with Hyperparameter Tuning on Models
    a. Origin: Assignment from DSC 550 Data Mining (Week 8)
    b. Description: The focus for this assignment was selection of the best model and hyperparameter tuning. The data set used was from Kaggle (G Ranjith Kumar) which centered around loan approvals. The high-level steps for this assignment were data preparation, creating a pipeline with min-max scaler and a KNN Classifier, Logistic Regression Classifier, and a Random Forest Classifier. A grid search was used to determine the best parameters for the various models and check which model performed the best.
7. MLB Attendance Improvement with Linear Regression
    a. Origin: Assignment from DSC 630 Predictive Analytics (Week 3)
    b. Description: The intent for this project was to use data from the Los Angeles Dodgers Baseball Team to determine how to improve attendance. The CRISP-DM methodology was followed for this analysis. The assignment portrays Exploratory Data Analysis (EDA), training/testing of a Linear Regression model, evaluation metrics to compare a model trained with reduced features, and recommended actions based on the insights gleaned.
8. ALS K-means Cluster Model
    a. Origin: Assignment from DSC 630 Predictive Analytics (Week 4)
    b. Description: The focus for this assignment was to use ALS patient data from PRO_ACT to apply clustering methods. The high-level highlights from this assignment consist of removing irrelevant data, applying a standard scalar to the data, plotting a cluster

silhouette score, fit a K-means model to the data, fit a PCA transformation with two features of the scaled data, generate a scatterplot of the PCA transformed data, and outlined a summary of the results.

9. Time Series Modeling with US Retail Sales Data
    a. Origin: Assignment from DSC 630 Predictive Analytics (Week 8)
    b. Description: As the title outlines, this assignment focused on time series analysis with US Retail Sales Data. The data consisted of total monthly retail sales in the US from January 1992 until June 2021. There were several plots of the data with respect to time. The data was split based on time (last year-target, previous years training) to prepare for training/testing the model. A Holt-Winters Forecast with Exponential Smoothing model was chosen for predictive forecasting. Root Mean Squared Error (RMSE) was the metric chosen to evaluate the performance of the model.
10. Recommender System for Movies
    a. Origin: Assignment from DSC 630 Predictive Analytics (Week 10)
    b. Description: The purpose of this assignment was to create a recommender system using the small MovieLens data set. Users could input a movie they like, and the recommender system will output 10 other recommendations to watch. The highlights from this assignment include Exploratory Data Analysis (EDA), consolidating a data frame based on correlation values, and providing an output summary of recommendations based on the user's movie input.
11. Digit and Image Classification with Neural Networks
    a. Origin: Assignment from DSC 650 Big Data (Week 6)
    b. Description: There were a few parts to this assignment, but the focus was classifying images with Neural Networks. The first part of the assignment focused on classifying digits from the MNIST digit data set. A CNN model was trained with the image data and evaluated for accuracy and loss plots. A second part of this assignment consisted of creating a CNN model to classify images of different objects using CIFAR10 small images classification data set. This was performed with and without data-augmentation. Lastly, the ResNet 50 model was used to perform image classification on images personally chosen for the assignment.
12. Variational Autoencoder with MNIST Data
    a. Origin: Assignment from DSC 650 Big Data (Week 12)
    b. Description: The purpose for this assignment was to implement a variational autoencoder with the MNIST data set. The high-level steps were creating the appropriate directories, creating a ConvNet, generating a sampling function, decoder implementation, creating a custom variational layer class, instantiating, and training the model, and then generating/saving a plot. A grid (15 x 15) of digits was saved as an output.

The two remaining projects from DSC 680 may be added to the portfolio as well. These additional projects have not been started at this time, so they were not included in this Milestone.