## **Project Proposal**

## **Student(s): Anthony Lam, Grace Pham, Quan Ta**

#### **Date: 08/16/2020**

#### **Purpose**

You are asked to propose a final project and present in the class. This proposal should describe the problem, the data sets, and the goal(s) of the project. Use the Project Requirements at the end of this notebook for choosing and scoping your project.

### **1. Introduction**

*(Introduce the project and describe the objectives.)*

Our project will be about predicting the amount of time that it will take to be approved for the permit based on the features that they have. This will be done by using the San Francisco Building Permit Data. This dataset is all the information that is collected that pertains to the permit issued by the government. These are all general information such as dates, type of permit, costs, and other kinds of needed information. Our objectives for this are to create a classification that can be used to predict the amount of days it takes. The kind of algorithm that we plan on using is a Naive Bayes algorithm to predict that kind of category that it will fall into. These categories will be: Same day, A week, 1-2 weeks, 2 weeks - 1 month, 1-2 months, 2-3 months, and so on. This will be done by using the features: permit type, estimated cost, revised cost, existing units, existing construction type, proposed construction type, and zip code. We will be measuring the accuracy of the project by measuring the accuracy through sklearn. We will also be trying out a KNN classifier to see how well that works. We hope to use a linear regression to be able to predict the days that it will take to get approved.

​

### **2. Problem Definition**

*(Define the problem that will be solved in this data analytics project.)*

The primary goal of this project is to predict the amount of time that it takes to be approved for the permit based on the features that were given. From visualizing and calculating the average wait time of the data, it took around 26 days to get a permit for most people. We felt that this was somewhat long for people and by being able to get an accurate estimate on just how long it would take, it can help create some pressure in getting your permit application status sped up. This can be very useful for someone who is looking to try to get a big project approved for a permit but they are worried that it may take too long to start their project. This could help in figuring out how long it would take based on the inputs that he’s giving and then allow him to plan ahead for this wait. This could also be beneficial for the ones approving the licenses as they’ll now have a better read on how long it’ll take to get through someone’s application and give them a better estimate as to how many days it’ll take them to issue a permit.

​

### **3. Data Sources**

*(Describe the origin of the data sources. What is the format of the original data? How to access the data?)*

The data source was acquired from Kaggle: <https://www.kaggle.com/aparnashastry/building-permit-applications-data>.

The initial purpose of the San Francisco Building Permit dataset was to record any permits for look-up in the future. However, the purpose of the dataset changes as the public believe that the delays in issuing building permits negatively influence the real estate industry. Therefore, this dataset was put together to investigate whether the building permits have anything related to the imbalance in demand and supply of the real estate market in San Francisco. We found this dataset on Kaggle. However, the dataset was originally constructed based on building permits that San Francisco authority issued and published on San Francisco open data portal. The format of the original dataset is a csv file and we can access this through creating a dataframe using Pandas.

​

### **4. The Goal(s) of the predictions**

*(What are the expected results of the project?)*

The goals of this prediction are to be able to accurately categorize the time that it will take to be approved for the permit. Through KNN and Naive Bayes, we are hoping to get at least 95% on accuracy, F1, precision, and recall scores. Another goal that we have is to also be able to predict the number of days that it will take to get your permit based on the time that you file and when it's issued. These will be affected by other parts of the features and then we can measure how well this model is doing based on the error in the model.

​

(*Use the following requirements for writing your reports. DO NOT DELETE THE CELLS BELLOW*)

# **Project Requirements**

This final project examines the level of knowledge the students have learned from the course. The following course outcomes will be checked against the content of the report:

Upon successful completion of this course, a student will be able to:

* Describe the key Python tools and libraries that related to a typical data analytics project.
* Identify data science libraries, frameworks, modules, and toolkits in Python that efficiently implement the most common data science algorithms and techniques.
* Apply latest Python techniques in data acquisition, transformation and predictive analytics for data science projects.
* Discuss the underlying principles and main characteristics of the most common methods and techniques for data analytics.
* Build data analytic and predictive models for real world data sets using existing Python libraries.

\*\* Marking will be foucsed on both presentation and content.\*\*

## **Written Presentation Requirements**

The report will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as its contents. Please make sure that the text of your report is well-structured, using paragraphs, full sentences, and other features of well-written presentation.

## **Technical Content:**

* Is the problem well defined and described thoroughly?
* Is the size and complexity of the data set used in this project comparable to that of the example data sets used in the lectures and assignments?
* Did the report describe the charactriatics of the data?
* Did the report describe the goals of the data analysis?
* Did the analysis conduct exploratory analyses on the data?
* Did the analysis build models of the data and evaluated the performance of the models?
* Overall, what is the rating of this project?

[ ]:

​