**Capstone Project**

**Final**

**Project proposal**

**1.** **Group description**

**1.1.** Group name

|  |
| --- |
| Partly Parrots x 7Park |

**1.2.** Students names, background and target industry if any

|  |
| --- |
| **Valeria (Val) Garland**  *Background*: BASc in Mineral Engineering from the University of Toronto ’18 (Canada), with 2 years of operational experience at a gold mine site in arctic Russia. Extensive experience in data analysis in Excel, with strong presentation skills to higher management.  *Target Industry:* Tech, but open to others  **Chun Tao**  *Background*: BS in Psychology (Peking University) and MS in Education Psychology (UPenn), with 5 years’ experience in business operations and marketing.  *Target Industry*: e-commerce & tech |

**1.3.** Group structure: roles and responsibilities

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Student** | **Data Science** | **Project Team** | | Val Garland | - Exploratory Data Analysis  - Data Cleaning  - Machine Learning  - Data Visualization (focus on plotly) | - GitHub repo owner  - Presentation | | Chun Tao | - Exploratory Data Analysis  - Data Cleaning  - Feature Engineering  - Machine Learning | - Presentation |   *Time Management* – deadlines will be set within both student’s calendars with reminders |

**2. Why** do we want to develop a data science project?

**2.1 Objective**: what problem do you want to solve? What questions are you trying to answer? How will you **measure the success** of your analysis from a business/user perspective?

|  |
| --- |
| **Context**  7Park Data approached NYCDSA to have its’ students look into forecasting the Zillow Rental Index from both public and private data sets. This will help 7Park gain valuable insights for their clients in the Commercial Real Estate sector, as well as explore new data – in our case this will be looking at severe climate events (forest fires, hurricanes, floods, etc.).  **Objective**   * Forecast the Zillow Rental Index (ZORI) * Highlight top 5 features in final predictive model * Look into how forest fires in California affect rental prices (NYCDSA objective) * Look into how other natural disasters affect rental price nationwide (7Park)   **Measure of Success**   * Back-testing to ZRI with high accuracy and low MSE |

**2.2.** **Scope** of application: what population and timeframe will your analysis/model be applied to or used for?

|  |
| --- |
| **Population**: process and model will be applied nationwide, based on US Census data.  **Timeframe**: Data from the past 7 years (going back to 2013) will be used from each data source (outlined in the following sections)  **Target variable**: median rental price for commercial real estate nationwide |

**3. How** do you translate the objective and scope in terms of data?

**3.1.** What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

|  |
| --- |
| *\*\*All of the data shown below is publicly available*  **Data 1: American Community Survey (ACS) (demographics)**   * Zip-code granularity * Yearly data   **Data 2: American Housing Survey**   * Yearly data   **Data 3: Zillow Rental Index (for historical Zillow data)**   * Zip-code granularity * Monthly data   **Data 4: IRS (income / unemployment data)**   * Zip-code granularity * Yearly data   **Data 5: Climate Data (TBD)** |

**3.2.** What **data treatment and analysis** do you plan? Data Aggregation, target variable definition, tools, analysis/machine learning, ...

|  |
| --- |
| **Data Preparation**   * Merge datasets for yearly IRS data * Merge datasets for yearly ACS data * Check missingness and deal with it appropriately   **Target Variable**   * Predict ZRI for 1 year * Apply criteria validated with stakeholders (discuss climate data potential)   **Tools**   * Google BigQuery for ACS and AHS data extraction * Downloading from IRS and Zillow for income + unemployment and ZRI data respectively * Data cleaning in python with pandas (data viz with matplotlib and plotly) * Python SKLearn for all machine learning models   **Analysis**   * Prediction model: test multiple linear regression as baseline and then look at adding climate data as another independent feature |

**4. Project Schedule**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **December ‘20** | | | | | | | | | | | | | | | |
|  | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| **7Park Clarification Call** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Exploration + Extraction  *Finding appropriate data* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Kick-Off with NYCDSA** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Cleaning |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ***Complete EDA by EOD*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ***Onboarding with 7Park HR*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ***Milestone 1 with 7Park*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ML: Multiple LR  *Base Model* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ML: Other models |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Climate Data Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ML with Climate Data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ***Milestone 2 with NYCDSA*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Presentation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Complete Presentation** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Project Submission** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Project Presentation** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

*\*The diagonally shaded cells indicate additional time* ***if*** *required*